

The background of the cover is a blue grid with various mathematical formulas in a light blue, semi-transparent font. These include binomial coefficients $\binom{n}{x}$, probability distributions like $p^x q^{n-x}$, and the exponential series $e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$.

Introduction to Regression Analysis

M.A. Golberg
H.A. Cho

 WITPRESS

Introduction to Regression Analysis

WIT*PRESS*

WIT Press publishes leading books in Science and Technology.

Visit our website for new and current list of titles.

www.witpress.com

WIT*eLibrary*

Making the latest research accessible, the WIT electronic-library features papers presented at Wessex
Institute of Technology's prestigious international conferences.

To access the library and view abstracts free of charge please visit www.witpress.com

This page intentionally left blank

Introduction to Regression Analysis

M.A. Golberg & H.A. Cho

Department of Mathematical Sciences
University of Nevada
Las Vegas, USA

٢٠٢٠



Introduction to Regression Analysis

M.A. Golberg & H.A. Cho

Published by

WIT Press

Ashurst Lodge, Ashurst, Southampton, SO40 7AA, UK

Tel: 44 (0) 238 029 3223; Fax: 44 (0) 238 029 2853

E-Mail: witpress@witpress.com

<http://www.witpress.com>

For USA, Canada and Mexico

WIT Press

25 Bridge Street, Billerica, MA 01821, USA

Tel: 978 667 5841; Fax: 978 667 7582

E-Mail: infousa@witpress.com

<http://www.witpress.com>

British Library Cataloguing-in-Publication Data

A Catalogue record for this book is available
from the British Library

ISBN: 1-85312-624-1

Library of Congress Catalog Card Number: 98-86352

*The texts of the papers in this volume were set
individually by the authors or under their supervision.
Only minor corrections to the text may have been carried
out by the publisher.*

No responsibility is assumed by the Publisher, the Editors and Authors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

© WIT Press 2004.

Revised and updated 2010

Reprinted 2010 by CPI Antony Rowe, Chippenham and Eastbourne, UK.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the Publisher.

Contents

Preface	xi
1 Introduction	1
1.1 A Brief History of Regression	1
1.1.1 Genealogy of Regression	1
1.1.2 The Method of Least Squares	2
1.2 Typical Applications of Regression Analysis	2
1.2.1 Use of Regression Analysis	2
1.2.2 Data Sets	2
1.3 Computer Usage	3
2 Some Basic Results in Probability and Statistics	5
2.1 Introduction	5
2.2 Probability Spaces	5
2.3 Random Variables	7
2.4 The Probability Distribution of X	7
2.5 Some Random Variables and their Distributions	8
2.6 Joint Probability Distributions	12
2.7 Expectation	15
2.7.1 Moments	16
2.7.2 Moment Generating Function	18
2.8 The Normal and Related Random Variables	21
2.8.1 Normal Random Variables	21
2.8.2 Chi-Square Random Variables	22
2.8.3 t and F -Distributions	23
2.8.4 Lognormal Random Variables	25
2.9 Statistical Estimation	25
2.9.1 The Method of Moments	25
2.9.2 Maximum Likelihood Estimation	28
2.9.3 Least Squares Estimation	31
2.9.4 Bayesian Estimation	31
2.10 Properties of Estimators	33
2.10.1 Consistency and Unbiasedness	33
2.10.2 Sufficiency	35
2.11 Confidence Intervals	37
2.11.1 Exact Confidence Intervals	37

2.11.2	Approximate Confidence Intervals	38
2.12	Hypothesis Testing	39
2.12.1	Best Tests	40
2.12.2	Generalized Likelihood Ratio Tests	43
2.13	Hypothesis Testing and Confidence Intervals	44
2.14	Exercises	45
3	Simple Linear Regression	51
3.1	Introduction	51
3.2	The Error Model	52
3.2.1	Algebraic Derivation of the Least Squares Estimators	56
3.3	Estimating σ^2	65
3.4	Properties of $(\hat{\beta}_0, \hat{\beta}_1, s^2)$	68
3.4.1	Standard Errors of the Coefficients	72
3.5	The Gauss-Markov Theorem	73
3.6	Confidence Intervals for (β_0, β_1)	76
3.6.1	Simultaneous Confidence Intervals	78
3.7	Hypothesis Tests for (β_0, β_1)	80
3.8	The ANOVA Approach to Testing	82
3.8.1	Regression Through the Origin	91
3.8.2	Estimation and Testing for Regression through the Origin	91
3.8.3	Prediction	95
3.9	Assessing Model Validity	101
3.9.1	The Lack of Fit Test (LOFT)	104
3.9.2	Residual Plots	111
3.10	Transformations	113
3.10.1	Transformations of x	117
3.10.2	Transformations in x and y	118
3.10.3	Box-Cox Transformations	119
3.11	Exercises	123
4	Random Vectors and Matrix Algebra	129
4.1	Introduction	129
4.2	Matrices and Vectors	129
4.2.1	Some Special Matrices	130
4.3	Fundamentals of Matrix Algebra	131
4.3.1	Matrix Addition	131
4.3.2	Properties of Matrix Addition	131
4.3.3	Matrix Multiplication	132
4.3.4	Properties of Matrix Multiplication	133
4.3.5	Scalar Multiplication	134
4.3.6	Powers of Matrices	134
4.3.7	Matrix Trace	135
4.4	Matrices and Linear Transformations	135
4.4.1	Matrix Inversion	136
4.4.2	The Inverse of Partitioned Matrices	139
4.4.3	The Sherman-Morrison-Woodbury Formula	141

4.5	The Geometry of Vectors	142
4.5.1	Length and Angle	144
4.5.2	Subspaces and Bases	146
4.6	Orthogonal Matrices	148
4.6.1	Eigenvectors and Eigenvalues	149
4.6.2	The Spectral Theorem for Symmetric Matrices	150
4.6.3	Some Further Applications of the Spectral Theorem	154
4.6.4	Expectation of Quadratic Forms	156
4.7	The Multivariate Normal Distribution	156
4.7.1	The Nondegenerate Case	156
4.7.2	The Degenerate Multivariate Normal Distribution	160
4.8	Solving Systems of Equations	161
4.8.1	Gaussian Elimination	162
4.8.2	Cholesky Factorization	166
4.9	The Singular Value Decomposition	167
4.9.1	The QR Decomposition	170
4.10	Exercises	171
5	Multiple Regression	179
5.1	Introduction	179
5.2	The General Linear Model	179
5.3	Least Squares Estimation	182
5.3.1	Estimating β	182
5.3.2	Some Analytical and Numerical Solutions of the Normal Equations	189
5.3.3	Numerical Examples	197
5.3.4	Estimating σ^2	206
5.4	Properties of $(\hat{\beta}, s^2, \hat{\varepsilon})$	206
5.4.1	Properties of $\hat{\varepsilon}$	211
5.4.2	Further properties of $\Sigma(\hat{\beta})$	212
5.4.3	A Summary of OLS Estimators	216
5.5	The Gauss-Markov Theorem	217
5.6	Testing the Fit - the Basic ANOVA Table	218
5.6.1	The Overall F -Test	218
5.6.2	The Coefficient of Multiple of Determination	224
5.7	Confidence Intervals and t -Tests for the Coefficients	226
5.7.1	Confidence Intervals	226
5.7.2	t -Tests	227
5.8	The Extra Sum of Squares Principle	232
5.8.1	The General Linear Hypothesis	232
5.8.2	The F -Test	234
5.8.3	Derivation of the F -Test	236
5.9	Prediction	240
5.9.1	Predicting $E(Y_{\mathbf{x}})$	240
5.9.2	Prediction Intervals	241
5.9.3	Extrapolation	242
5.10	Exercises	243

6	Residuals, Diagnostics and Transformations	249
6.1	Introduction	249
6.2	Residuals	250
6.2.1	Properties of $\hat{\varepsilon}$	250
6.2.2	The Leverage h_{ii}	251
6.3	Residual Plots	252
6.3.1	Normal Plots	252
6.3.2	Variable Plots	252
6.3.3	Partial Plots	253
6.4	PRESS Residuals	265
6.4.1	Deletion Statistics	266
6.4.2	Influence Diagnostics	271
6.4.3	Influence on $\hat{\beta}$, \hat{y}_i	271
6.5	Transformations	276
6.5.1	Transformations in \mathbf{x}	276
6.5.2	The Box-Tidwell Method	277
6.5.3	Transformations of \mathbf{y}	279
6.5.4	Linearizable Transformations	280
6.5.5	Box-Cox Transformations	280
6.5.6	Quick Estimates of λ	281
6.5.7	Variance Equalizing Transformations	288
6.5.8	Variance Stabilizing Transformations	299
6.6	Correlated Errors	301
6.6.1	The Durbin-Watson Statistic	301
6.6.2	Correcting for Autocorrelation	302
6.7	Generalized Least Squares	305
6.8	Exercises	308
7	Further Applications of Regression Techniques	313
7.1	Introduction	313
7.2	Polynomial Models in One Variable	313
7.2.1	Orthogonal Polynomials	315
7.2.2	Piecewise Polynomial Models	317
7.2.3	Multivariate Polynomial Models	322
7.3	Radial Basis Functions	323
7.3.1	Types of Radial Basis Functions	323
7.3.2	Fitting Methods for RBFs	325
7.4	Dummy Variables	327
7.4.1	Further Comments on Dummy Variable	334
7.5	Interactions	337
7.6	Logistic Regression Revisited	341
7.6.1	Interpretation of Logistic Coefficients	345
7.6.2	Maximum Likelihood Estimation	347
7.7	The Generalized Linear Model	348
7.7.1	Linear Predictors and Link Functions	349
7.7.2	The Error Function	350
7.7.3	Parameter Estimation	351
7.8	Exercises	352

8	Selection of a Regression Model	359
8.1	Introduction	359
8.2	Consequences of Model Misspecification	360
8.3	Criteria Functions	361
8.3.1	Coefficient of Multiple Determination R^2	362
8.3.2	Mallows' C_p	363
8.3.3	The PRESS Statistic	366
8.3.4	Standardized Residual Sum of Squares	367
8.3.5	Other Criteria	367
8.4	Various Methods for Model Selection	367
8.4.1	Evaluating All Possible Regressions	368
8.4.2	Backward Elimination	369
8.4.3	Forward Selection	372
8.4.4	The Stepwise Regression Procedure	373
8.4.5	Selection of Models - An Overview	375
8.5	Exercises	376
9	Multicollinearity: Diagnosis and Remedies	379
9.1	Introduction	379
9.2	Detecting Multicollinearity	380
9.3	Other Multicollinearity Diagnostics	384
9.3.1	Consequences of Multicollinearity	387
9.3.2	Prediction	388
9.4	Combatting Multicollinearity	389
9.5	Biased Estimation	390
9.5.1	Shrunken Estimators	390
9.5.2	Ridge Regression	392
9.5.3	Choosing the Ridge Parameter	397
9.5.4	Generalized Ridge Regression	404
9.6	Other Alternatives to OLS	407
9.6.1	Mixed Estimation	407
9.6.2	Principal Components Estimation	408
9.7	Exercises	409
	Appendix	413
	Bibliography	421
	Index	429

This page intentionally left blank

Preface

Regression analysis has been one of the most widely used statistical methodologies during the past 50 years for analyzing relationships among variables. Due to its flexibility, usefulness, applicability, theoretical and technical succinctness, regression analysis has become a basic statistical tool to solve problems in the real world. In order to apply these elegant techniques successfully and effectively, one requires sound insight and understanding of both the underlying theory (i.e., statistical reasoning) and its practical application.

This book is designed primarily as a standard course in regression analysis, and is an outgrowth of class notes for advanced undergraduates, graduate students and researchers in various fields of engineering, the chemical and physical sciences, mathematical sciences and statistics. Therefore it blends both theory and application so that the reader will gain a deep enough understanding of the basic principles necessary to apply regression model building techniques in a wide variety of environments. It contains conventional topics and recent practical developments.

This book is also intended to fill a gap in what we perceive to be a communication gap faced by students or researchers who have a limited mathematical background but would like to continue from where most beginning statistics texts end, or who want to build up their knowledge of advanced statistical data analysis and modeling.

The book consists of nine chapters. The first seven chapters are devoted to a fairly comprehensive description of linear regression modeling, methods of analysis and their ramifications. Chapters 8 and 9 are the concluding chapters which present the decision-making aspects of this book.

In Chapter 1 we give a general introduction to the scope of regression analysis, its brief historical background and describe some typical applications of regression.

Chapter 2 provides a review of some basic results in probability theory and basic concepts in statistics, which are essential to understand the general framework of the statistical method and to develop the further theory and its applications.

In Chapter 3 we introduce the method of least squares for fitting straight lines. Since simple linear regression is the gateway to the analysis of the general linear model, we provide detailed theoretical aspects and a broad applied viewpoint as well.

Chapter 4 provides a rigorous outline of indispensable results in linear algebra including matrix theory. This will help readers facilitate their understanding of later chapters. Matrix notation that is used throughout the book is introduced to improve readers' confidence and understanding.

Chapter 5 discusses the study of multiple regression analysis along with a comprehensive approach to inference procedures. The material in Chapter 5 constitutes the core of the text.

In Chapter 6, we are concerned with the analysis of residuals and detecting violations from model assumptions. Modern diagnostic methods are also discussed for outlier detection, plotting of residuals and the development of transformation techniques.

Chapter 7 then considers more complicated models in linear regression and their applications. These include polynomial models, radial basis functions, the use of dummy variables, logistic regression for which the response is qualitative/binary, and a basic treatment of the generalized linear model.

Chapter 8 returns to the topic of evaluation criteria for subset selection and discusses various types of selection procedures.

Finally, Chapter 9 is devoted to diagnosis and correction of multicollinearity which are common and serious problems in regression analysis.

For classroom purposes, it is suggested that two semesters be taken for complete coverage. However it is possible to treat most of topics in one semester if some of the pre-requisite knowledge is assumed or to make the level/depth of the treatment less rigorous. Exercises are provided at the end of chapters two to nine. We have tried to balance theoretical aspects and applications to data analysis. The data are mostly based on real world situations.

In addition, even though this book is not intended as a manual for any statistical computer package, readers can use computer packages to apply the techniques learned here. However, it is strongly recommended that one use high quality software only after readers thoroughly understand the statistical reasoning for the methods being used.

The authors wish to give warmest thanks to Professor Carlos Brebbia who gave us the opportunity to get this material into print and to Mr. Brian Privett, head of production of WIT who has been waiting patiently for the manuscript's completion. We would like to thank Professor C.S. Chen for introducing the authors. We would also like to thank a number of friends, neighbors and colleagues in the Department of Mathematical Sciences at the University of Nevada, Las Vegas. Finally, we are grateful to our wives, Joyce and Sookhyun Ellen, who encouraged us to complete this project under difficult circumstances.

Michael A. Golberg
Hokwon A. Cho

To our wives

Joyce
Sookhyun

To our children

Jonathan
Stefany
Katherine Soojin

This page intentionally left blank

Chapter 1

Introduction

Regression analysis is a collection of statistical techniques that serve as a basis for drawing inferences about relationships among interrelated variables. Since these techniques are applicable in almost every field of study, including the social, physical and biological sciences, business and engineering, regression analysis is now perhaps the most used of all data analysis methods. Hence, the goal of this text is to develop the basic theory of this important statistical method and to illustrate the theory with a variety of examples chosen from economics, demography, engineering and biology. To make the text relatively self contained we have included basic material from statistics, linear algebra and numerical analysis. In addition, in contrast to other books on this topic [27, 87], we have attempted to provide details of the theory rather than just presenting computational and interpretive aspects.

1.1 A Brief History of Regression

1.1.1 Genealogy of Regression

A well-known British anthropologist Sir Francis Galton (1822-1911) seems to be the first to introduce the word “regression” in his study on heredity. He found that on the average, heights of children do not tend toward the parents’ heights, but rather toward the average as compared to the parents. Galton termed this “*regression to mediocrity in hereditary stature*.” In the *Journal of the Anthropological Institute*, Vol. 15 (1885), pp. 246-263, it says that “... The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it.” Galton then described how to determine the relationship between childrens’ heights using parents’ heights. Today Galton’s analysis would be called a “*correlation analysis*,” a term for which he is also responsible. In most model-fitting situations today, there are no elements of “*regression*” in the original sense. Nevertheless, the word is so established that we continue to use it. For more related stories about the history of regression, we refer readers to the Statistical Encyclopedia or The History of Statistics by Stigler (1986) [110].

1.1.2 The Method of Least Squares

As we shall see, the basic mathematical tool in regression analysis is the *method of least squares*. There has been a controversy concerning who first discovered the method of least squares. Generally, credit for the method of least squares is given to Carl Friedrich Gauss (1777-1855). Apparently Adrien Marie Legendre (1752-1833) seemed to work independently on its use in 1805. In brief, the idea of the *least squares method* is to find the values of the unknown constants in a hypothesized equation that minimizes the sum of the squared deviations of the observed values from those predicted by the model. The justification for this is given by the celebrated *Gauss-Markov theorem*, which is proved in Chapters 3 and 5.

However, the thing we need to focus on is that regression analysis and the method of least squares have always been linked to practical use.

1.2 Typical Applications of Regression Analysis

Since regression analysis is not just fitting equations to data, the crucial point is that for every problem in various fields of science one needs to clarify the goal of the problem and the methods needed for the regression analysis. The following is a summary to provide some guidelines.

1.2.1 Use of Regression Analysis

The main purposes of regression analysis can be summarized as:

1. Data description - to investigate or refute a relationship among variables.
2. Interpretation - to give a summary or an interpretation through the fitted model to obtain an interpolation or calibration curve/surface.
3. Inference - to develop or improve the theoretical model(s)/method(s) which should be chosen to extend and generalize it to other sets of data. These formal statistical techniques are called *estimation of parameters*, *testing* and *prediction*.

We assume that a set of data has been obtained. For practical application of the methodology, it is important to remember that regression analysis is a data-analysis oriented approach to problem solving. Hence, fitting an equation itself may not be the primary objective of the study. Furthermore, fitting an equation may only be an intermediate process to gain insight and understanding of the data.

1.2.2 Data Sets

An essential part of regression analysis is the data which have been collected. Perhaps the most serious limitation in a regression analysis is to fail to collect data on all potentially important regressors. It is fair to say that the results are only as good as the data that produced them. The basic methods of collecting data are:

1. Retrospective study - using historical data or an existing census, etc.
2. Observational study - through random observations.

3. Experimental study - through designed/planned experiments or surveys.

Even if many of the difficulties due to data sets are rather obvious, limitations in the data gathering process may prevent them from being analyzed appropriately.

1.3 Computer Usage

During the past twenty years advances in computer technology particularly in personal computers (PCs), and telecommunications such as the internet (World-Wide Web) have brought us revolutionary improvements in statistical research. Because of these developments many of the traditional limitations in the computational aspects of regression analysis have largely disappeared. In consequence a large number of high quality statistical packages have become available which have largely eliminated the drudgery of regression calculations so that students, practitioners and researchers can focus on the analysis of data rather than just calculating descriptive aspects.

As a result, many new mathematical techniques have been developed for data analysis which would not have been feasible 20 years ago. In addition, most currently popular packages such as MINITAB[®], SAS[®] and SPSS[®] are now available for PCs with user-friendly spreadsheet interfaces which make calculations easy to do. Of particular importance is the ease in which high quality graphical output can be obtained. Of course such tools should not be used uncritically, and it is one of the major goals of this text to make it possible for students to use these packages with a certain confidence that they understand their output. Because of its widespread availability, quality and ease of use we have chosen MINITAB[®] to do most of the calculations and plots that appear in this book.

Moreover, the internet has made it possible to obtain a large variety of additional material, such as data sets, lecture notes and free computing software which are useful adjuncts to the material presented here. Much of this can be obtained with a few simple mouse clicks using currently available search engines. We strongly advise students to make use of these capabilities.

Chapter 2

Some Basic Results in Probability and Statistics

2.1 Introduction

Throughout this text we will assume that the student has had a basic course in probability theory and statistical inference such as that in Refs [40, 63]. Our preference is that such a course would be at a level requiring calculus and some background in matrix theory. Because some tools in matrix algebra are necessary for multiple regression, we will outline the necessary ones in Chapter 4. In this chapter we will present some of the standard results in probability and statistics that will be needed throughout the book. Those familiar with this material may skip this chapter and simply refer to it as necessary.

2.2 Probability Spaces

A *probability space* (also called a *sample space*) is a set Ω together with a collection of subsets A of Ω called *events*. If A is an event, the probability that A occurs is written as $P(A)$ and for short, is usually read as the “probability of the event A .” As is well known, $P(A)$ has the following properties:

- (1) $0 \leq P(A) \leq 1$;
- (2) $P(\Omega) = 1$;
- (3) $P(\phi) = 0$, where ϕ is the empty or impossible event;
- (4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where $A \cup B$ is the event that “either A or B occurs” and $A \cap B$ is the event that “both A and B occur.” If A and B cannot occur simultaneously ($A \cap B = \phi$), then $P(A \cap B) = 0$ and $P(A \cup B)$ reduces to

$$P(A \cup B) = P(A) + P(B). \quad (2.1)$$

In this case we say that A and B are *mutually exclusive events*.

Similarly, more complicated rules exist for calculating $P(A_1 \cup A_2 \cup \dots \cup A_n)$ which may be found in [40, 63]. For example, if $A_i, i = 1, 2, \dots, n$ are mutually exclusive; i.e.,

$$A_i \cap A_j = \phi, i \neq j, \quad (2.2)$$

then

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{j=1}^n P(A_j). \quad (2.3)$$

- (5) The event \bar{A} denotes the *complement* of A and is equivalent to saying that “ A does not occur.” Then it follows from (2.1) that

$$P(\bar{A}) = 1 - P(A). \quad (2.4)$$

- (6) If an event B is known to have occurred, then we may regard B as our new sample space and then “the probability of A given that B has occurred” (then both A and B have occurred) is called the *conditional probability of A given B* and is denoted by $P(A|B)$ and is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0. \quad (2.5)$$

- (7) Let the sample space be partitioned into k mutually exclusive events B_1, B_2, \dots, B_k such that $P(B_j) > 0, j = 1, 2, \dots, k$. Also let A be another event such that $P(A) > 0$, that has occurred in the sample space. Then

$$\begin{aligned} A &= A \cap (B_1 \cup B_2 \cup \dots \cup B_k) \\ &= (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_k). \end{aligned} \quad (2.6a)$$

However, $P(A \cap B_j) = P(B_j) P(A|B_j), j = 1, 2, \dots, k$ using (2.5). So

$$\begin{aligned} P(A) &= P(B_1) P(A|B_1) + P(B_2) P(A|B_2) + \dots + P(B_k) P(A|B_k) \\ &= \sum_{j=1}^k P(B_j) P(A|B_j). \end{aligned} \quad (2.6b)$$

This is called the *law of total probability*. Using (2.5) and (2.6b) we have

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i) P(B_i)}{\sum_{j=1}^k P(B_j) P(A|B_j)},$$

which is known as *Bayes' theorem*. The conditional probability $P(B_i|A)$ is expressed as a function of the simple probability of $B_i, P(B_i)$, which in the absence of information about A , is called the *prior probability*. On the other hand, $P(B_i|A)$ is called the *posterior probability*. In particular, for $k = 2$, this reduces to

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)}.$$

As we see, this describes how to revise the prior probability of the event B_1 in the light of additional information to yield the posterior probability.

If the occurrence of B has no effect on the occurrence of A then, $P(A|B) = P(A)$ and then (2.5) becomes

$$P(A \cap B) = P(A)P(B). \quad (2.7)$$

In this case we say that A and B are *independent events*. For more than two events the conditions for independence are more complicated. For example, for three events A, B, C , they are independent if and only if (iff)

$$P(A \cap B) = P(A)P(B), P(A \cap C) = P(A)P(C), P(B \cap C) = P(B)P(C), \quad (2.8a)$$

and

$$P(A \cap B \cap C) = P(A)P(B)P(C). \quad (2.8b)$$

It is important to note that (2.8a) does not imply (2.8b) nor does (2.8b) imply (2.8a). For n events $A_i, 1 \leq i \leq n$, they are independent if and only if for any subset of k events $A_{i_j}, 2 \leq j \leq k$ that

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}). \quad (2.9)$$

In particular,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n). \quad (2.10)$$

Equation (2.10) will be used repeatedly throughout the text.

2.3 Random Variables

Without getting into technical details, a *random variable* may be regarded as a real-valued function on a probability space. We will distinguish between two types of random variables (often abbreviated as r.v.) discrete and continuous. A *discrete random variable* is one which takes at most a denumerable number of values (often, just finitely many) while a *continuous random variable* may take on a continuum of values.

The set of values a random variable can take on is called its *range*. Typically, we will denote a random variable by upper case letters X, Y, Z etc. and the range of X will be denoted by $\mathbb{R}(X)$. For example, if we roll a die and X denotes the number of spots observed on top, then X is discrete and $\mathbb{R}(X) = \{1, 2, 3, 4, 5, 6\}$. On the other hand, if X denotes the value of a number chosen at random from the interval $[0, 1]$, then $\mathbb{R}(X) = [0, 1]$ and X is a continuous random variable.

Often in practice, if a random variable which is truly discrete, but can take on a very large number of values, it is customary to use a continuous random variable to approximately model the discrete random variable. For example, if we consider the amount of money a person can earn in a given year (in dollars), then strictly speaking $\mathbb{R}(X) = \{0, 1, 2, \dots\}$. However, it is often useful to approximate this random variable by one that can take on all values in $[0, \infty)$. This is quite common in statistical analysis and is usually done without comment where necessary. This convention will be followed throughout the text.

2.4 The Probability Distribution of X

If a random variable is discrete and $x_i \in \mathbb{R}(X)$, then

$$f_X(x_i) = P\{X = x_i\}, 1 \leq i \leq \infty, \quad (2.11)$$

denotes the probability that X takes on the value x_i . The sequence of values $\{f_X(x_1), f_X(x_2), \dots, f_X(x_n), \dots\}$ is called the *distribution* of X and is usually abbreviated as

$$f_X(x), \quad x \in \mathbb{R}(X). \quad (2.12)$$

Often, f_X is referred to as the *probability mass function* of X .

Example 2.1 Consider an experiment of tossing a fair coin twice. Let X be the number of heads observed. The sample space is $S = \{HH, HT, TH, TT\}$, then X can be either 0 or 1 or 2. Assuming the tosses are independent, we have the probability mass function of X as follows;

x	0	1	2
$P(X = x)$	1/4	1/2	1/4

More generally, the probability that X takes on a value between a and b , is denoted by

$$P\{a \leq X \leq b\} \quad (2.13)$$

and is given by

$$P\{a \leq X \leq b\} = \sum_{a \leq x \leq b} f_X(x) \quad (2.14)$$

where $f_X(x) = 0$ if $x \notin \mathbb{R}(X)$ so the sum in (2.14) has at most countably many values.

When X is a continuous random variable it will usually be the case (at least in this book) that

$$P\{a \leq X \leq b\} = \int_a^b f_X(x) dx \quad (2.15)$$

where $f_X(x)$ in this case is called the *probability density function* (pdf) of X .

In particular,

$$P\{X \leq x\} = \int_{-\infty}^x f_X(x) dx \equiv F_X(x), \quad (2.16)$$

is called the *cumulative distribution function* (cdf) of X (*distribution function* for short). By the fundamental theorem of calculus

$$f_X(x) = \frac{d}{dx} F_X(x). \quad (2.17)$$

2.5 Some Random Variables and their Distributions

For future reference we give a number of examples of commonly occurring random variables and their distributions. Further examples will be given in Section 2.8.

Example 2.2 (Bernoulli random variables) Many experiments can be described by observing one of two possible outcomes. For example, a coin toss can result in a “heads” or “tails”, a poll question is answered “yes” or “no” a medical treatment can be a “success” or “failure”. In this case we can take our sample space $\Omega = \{S, F\}$ and define a random variable X which counts the number of “successes.” Then $X(S) = 1$

and $X(F) = 0$. If the probability of success is p , then $P\{X = 1\} = f_X(1) = p$ and $P\{X = 0\} = f_X(0) = 1 - p = q$. Then $\mathbb{R}(X) = \{0, 1\}$ and

$$f_X(x) = p^x q^{1-x}, \quad x = 0, 1 \quad (2.18)$$

is the distribution of X . A random variable having the distribution in (2.18) is called a *Bernoulli random variable*.

Example 2.3 (Binomial random variables) If the experiment in Example 2.2 is repeated independently n times, then the sample space can be described by the Cartesian product $\Omega^n = \{S, F\}^n$ of Ω taken n times. Ω^n is the set of all ordered n -tuples of S 's and F 's and has 2^n elements. If (x_1, x_2, \dots, x_n) , $x_i = S$ or F , $1 \leq i \leq n$, is an element of Ω^n , then by independence $P\{(x_1, x_2, \dots, x_n)\} = p^k q^{n-k}$ where k is the number of successes in the sequence (x_1, x_2, \dots, x_n) . Now let $X : \Omega \rightarrow \mathbb{R}$ be the random variable which counts the number of successes in (x_1, x_2, \dots, x_n) , then $\mathbb{R}(X) = \{0, 1, 2, \dots, n\}$ and standard counting arguments show that the distribution of X is given by

$$f_X(x) = \binom{n}{x} p^x q^{n-x}, \quad 0 \leq x \leq n, \quad (2.19)$$

where

$$\binom{n}{x} = \frac{n!}{(n-x)!x!} \quad (2.20)$$

is called the *binomial coefficient*. (For n a positive integer, $n! = n(n-1) \cdots 2 \cdot 1$ and $0! = 1$.)

The distribution (2.19) is usually called the *binomial distribution* and X is called a *binomial random variable*.

Example 2.4 (Poisson random variables) As we mentioned previously, even when a random variable has a finite range, it is often convenient mathematically to approximate it by a random variable with an infinite range. An important example of this occurs when X is a binomial random variable when p is small and n is large. In this case, if we let $\lambda = np$, then (2.19) can be approximated by

$$\frac{n!}{(n-x)!x!} p^x q^{n-x} \simeq \frac{e^{-\lambda} \lambda^x}{x!}. \quad (2.21)$$

If we now let X take on values $\{0, 1, 2, \dots\} = \Omega$ and define $X : \Omega \rightarrow \mathbb{R}$ by $X(x) = x$ and

$$f_X(x) = P\{X = x\} = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x \geq 0, \quad (2.22)$$

then $f_X(x) \geq 0$ and

$$\sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1 \quad (2.23)$$

so that $f_X(x)$ is a probability distribution on Ω . This distribution is usually called a *Poisson distribution* and the corresponding random variable a *Poisson random variable*. The Poisson distribution is often called the *distribution of rare events* and is often used in statistical analyses to describe the occurrence of events occurring randomly in time.

Example 2.5 (Uniform random variables) To model an experiment where there are n equally probable numerical outcomes we let $\Omega = \{x_1, x_2, \dots, x_n\}$ where x_i is the i -th outcome, $1 \leq i \leq n$. Let $X : \Omega \rightarrow \mathbb{R}$ be defined by $X : \Omega \rightarrow \mathbb{R}$ by $X(x) = x$ gives the value of the i -th outcome, then

$$f_X(x_i) = P\{X = x_i\} = \frac{1}{n}, \quad 1 \leq i \leq n, \quad (2.24)$$

and X is called a *uniform random variable* and the corresponding distribution a *uniform distribution*.

If there are an infinite number of equiprobable outcomes, then one cannot use a discrete random variable to model such situations. In this situation a reasonable probabilistic model is to assume that the outcomes can occur in a finite interval of real numbers $-\infty < a < b < \infty$. In this case we take $\Omega = [a, b]$ and define $X : \Omega \rightarrow \mathbb{R}$ by $X(x) = x$, to describe the outcome of a point x chosen at random from $[a, b]$. Of course X is not a discrete random variable so we cannot use $f_X(x) = P\{X = x\}$ to describe its probabilistic properties.

Here, we need to modify our approach to assigning probability measures - we cannot begin by assigning probabilities to points, rather we must begin by assigning measures to intervals. To model the notion of an equiprobable choice of "points", intervals of equal length should be equally likely to occur. Thus, if $A = [c, d] \subseteq [a, b]$ we define

$$P(A) = \frac{d - c}{b - a}. \quad (2.25)$$

Thus, $P(A) \geq 0$ and $P(\Omega) = (b - a) / (b - a) = 1$. The measure defined by (2.25) is called the *uniform measure* on $[a, b]$ and then

$$P(c \leq X \leq d) = P\{[c, d]\} = \frac{d - c}{b - a}. \quad (2.26)$$

From (2.26) it follows that the distribution function of X is given by

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x - a}{b - a}, & a \leq x \leq b \\ 1, & x > b. \end{cases} \quad (2.27)$$

Differentiation of (2.27) gives

$$f_X(x) = \frac{d}{dx} F_X(x) = \begin{cases} 0, & x < a \\ \frac{1}{b - a}, & a \leq x \leq b \\ 0, & x > b, \end{cases} \quad (2.28)$$

$f_X(x)$ is called a *uniform density*.

Example 2.6 (Canonical random variables) It is important to observe from Example 2.5 that the distribution of a random variable X is inherited from the measure that is imposed on the underlying probability space. Hence, given a continuous function $F(x) : \mathbb{R} \rightarrow \mathbb{R}$ having the properties:

(i) $0 \leq F(x) \leq 1$;

(ii) $F(x)$ is a nondecreasing function, i.e., $F(x) \leq F(y)$ for $x < y$;

then it can be used to define a probability measure on \mathbb{R} by defining

$$P(a < X \leq b) = F(b) - F(a). \quad (2.29)$$

If we define the random variable $X : \mathbb{R} \rightarrow \mathbb{R}$ by $X(x) = x$, then it follows from (2.29) that the cdf of X is given by

$$F_X(x) = P\{X \leq x\} = F(x). \quad (2.30)$$

Thus, given a continuous function satisfying (2.30) we can always define a random variable X having the cdf $F(x)$. This random variable is called the *canonical random variable* associated with a given cdf $F(x)$. This justifies the common practice in probability theory and statistics of referring to random variables and distributions interchangeably. We shall follow this custom throughout the text.

If $F(x)$ is differentiable, then

$$f(x) = \frac{dF(x)}{dx}, \quad (2.31)$$

is called the *density* of $F(x)$. Again, by the fundamental theorem of calculus,

$$F(x) = \int_{-\infty}^x f(x) dx. \quad (2.32)$$

Thus a random variable may be expressed in terms of its density since the density of the cdf of a canonical random variable X satisfies

$$f_X(x) = f(x). \quad (2.33)$$

Hence, we may refer to a continuous random variable by its density and we shall do so without further comment.

Example 2.7 (Logistic random variables) Let

$$f(x) = \frac{e^x}{(1 + e^x)^2}, \quad -\infty < x < \infty, \quad (2.34)$$

then it is easily verified by making the change of variable $u = e^x$ that

$$\int_{-\infty}^{\infty} f(x) dx = 1. \quad (2.35)$$

Since $f(x) \geq 0$, $f(x)$ is a density function with cdf

$$F(x) = \frac{e^x}{1 + e^x}, \quad -\infty < x < \infty. \quad (2.36)$$

The density function given by (2.34) is called a *logistic density* and a random variable having a logistic density is called a *logistic random variable*.

It is often convenient in probability theory and statistics to be able to classify random variables whose densities are of seemingly different forms into families which have a common structure. An important family in modern statistics which contains the binomial, Poisson and normal random variables, are the *exponential densities* which have the form

$$f(x; \theta) = \exp [a(x) b(\theta) + c(\theta) + d(x)] \quad (2.37)$$

whose x can take on a discrete or continuous range of values independent of θ . This family gives rise to the generalized linear model of statistical models which contain as a particular case, the normal models, which will be the primary focus of this text.

Example 2.8 (Exponential family of random variables) As particular cases of (2.37) we show that Poisson and binomial random variables are members of the *exponential family*. In Section 2.8 we show that normal random variables are as well.

If X is a Poisson random variable, then its density is given by $f(x) = \exp(-\lambda) \lambda^x / x!$. Now $\lambda^x = \exp(x \log \lambda)$ so that $\exp(-\lambda) \lambda^x / x! = \exp(x \log \lambda - \lambda - \log x!)$. Thus, if $\theta = \lambda$, $a(x) = x$, $b(\theta) = \log \theta$, $c(\theta) = -\theta$, $d(x) = -\log x!$, $f(x; \theta)$ is of the form in (2.37) so a Poisson density belongs to the exponential family.

Similarly, for a binomial density one has

$$\begin{aligned} f(x; p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \exp \left[x \log p - x \log (1-p) + n \log (1-p) + \log \binom{n}{x} \right]. \end{aligned}$$

So, with $\theta = p$, $a(x) = x$, $b(\theta) = \log [\theta / (1-\theta)]$, $c(\theta) = n \log (1-p)$ and $d(x) = \log \binom{n}{x}$, $f(x; p)$ is of the form in (2.37) so is a member of the exponential family.

2.6 Joint Probability Distributions

Often we will consider many random variables simultaneously. The joint probabilistic behavior of n random variables $X_i, 1 \leq i \leq n$, is generally described by their *joint distribution function*

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, \dots, X_n \leq x_n\} \quad (2.38)$$

where the comma in (2.38) indicates intersection and \mathbf{X} is shorthand for the n random variables $X_i, 1 \leq i \leq n$. If all the random variables are continuous (strictly speaking *absolutely continuous*) then $F_{\mathbf{X}}(x_1, x_2, \dots, x_n)$ can be given as a multiple integral of their joint density $f_{\mathbf{X}}(x_1, x_2, \dots, x_n)$. For example, if X_1 and X_2 are two continuous random variables, then

$$F_{\mathbf{X}}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{\mathbf{X}}(t_1, t_2) dt_1 dt_2. \quad (2.39)$$

As for (2.39) one can recover $f_{\mathbf{X}}$ from $F_{\mathbf{X}}$ by differentiation; i.e.,

$$f_{\mathbf{X}}(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{\mathbf{X}}(x_1, x_2) \quad (2.40)$$

and more generally

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{\mathbf{X}}(x_1, x_2, \dots, x_n). \quad (2.41)$$

If the X 's are discrete, then $F_{\mathbf{X}}$ is given by summation of the *joint discrete density* $f_{\mathbf{X}}(x_1, x_2, \dots, x_n)$. For example if X_1 and X_2 are two discrete random variables, then

$$F_{\mathbf{X}}(x_1, x_2) = \sum_{y_2 \leq x_2} \sum_{y_1 \leq x_1} f_{\mathbf{X}}(y_1, y_2). \quad (2.42)$$

In the discrete case one can recover $F_{\mathbf{X}}$ from $f_{\mathbf{X}}$, but the process involves technical limiting arguments, which fortunately we will not need. In general, one usually specifies the joint behavior of continuous and discrete random variables by specifying their joint densities; the joint distributions $F_{\mathbf{X}}$ are then obtained by integration or summation.

If one knows the joint distribution or joint density of n random variables, then the joint distribution or joint density of any subset of these can be obtained by partial integration or summation. For example, if X_1 and X_2 are two continuous random variables, then the density of X_1 is given by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2) dx_2 \quad (2.43)$$

and the density of X_2 by

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2) dx_1. \quad (2.44)$$

If X_1 and X_2 are discrete random variables, then

$$f_{X_1}(x_1) = \sum_{x_2 \in \mathbb{R}(X_2)} f_{\mathbf{X}}(x_1, x_2) \quad (2.45)$$

and

$$f_{X_2}(x_2) = \sum_{x_1 \in \mathbb{R}(X_1)} f_{\mathbf{X}}(x_1, x_2). \quad (2.46)$$

The distributions obtained by summation or integration of the joint distributions are usually referred to as the *marginal distributions* or densities of the joint distribution and by definition these are known once the joint distributions are known. In general, one cannot reverse this process. That is, knowing the marginal distributions of n random variables, one cannot reconstruct the joint distribution or densities. Put another way, there may be many joint distributions which have the same marginal distributions.

Example 2.9 Let X_1 and X_2 have the joint pdf

$$f(x_1, x_2) = \begin{cases} 10x_1x_2^2, & 0 < x_1 < x_2 < 1 \\ 0, & \text{elsewhere.} \end{cases} \quad (2.47)$$

The marginal pdf of X_1 is

$$f_{X_1}(x_1) = \int_{x_1}^1 10x_1x_2^2 dx_2 = \frac{10}{3}x_1(1 - x_1^3), \quad 0 < x_1 < 1, \quad (2.48)$$

zero elsewhere, and the marginal pdf of X_2 is

$$f_{X_2}(x_2) = \int_0^{x_2} 10x_1x_2^2dx_1 = 5x_2^4, \quad 0 < x_2 < 1, \quad (2.49)$$

zero elsewhere. Furthermore, the *conditional density* of X_2 given $X_1 = x_1$, $f(x_2|x_1)$ is given by

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{10x_1x_2^2}{\frac{10}{3}x_1(1-x_1^3)} = \frac{3x_2^2}{(1-x_1^3)}, \quad 0 < x_1 < x_2 < 1. \quad (2.50)$$

In contrast to Example 2.9 there is an important situation where the marginal distributions determine the joint distribution and that is when the random variables are independent.

Definition 2.1 Let $X_i, 1 \leq i \leq n$, be random variables. We say that $X_i, 1 \leq i \leq n$, are *independent random variables* if and only if the joint distribution $F_{\mathbf{X}}(x_1, x_2, \dots, x_n)$ factors as a product of n distributions $G_{X_i}(x_i), 1 \leq i \leq n$. That is,

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n G_{X_i}(x_i), \quad (2.51)$$

where $G_{X_i}(x_i), 1 \leq i \leq n$, are distribution functions. That is,

$$0 \leq G_{X_i}(x_i) \leq 1, \quad 1 \leq i \leq n \quad (2.52)$$

and

$$\lim_{x \rightarrow -\infty} G_{X_i}(x_i) = 0 \text{ and } \lim_{x \rightarrow \infty} G_{X_i}(x_i) = 1. \quad (2.53)$$

From Definition 2.1 it follows that if the factorization in (2.51) holds, then $G_{X_i}(x_i) = F_{X_i}(x_i), 1 \leq i \leq n$, are the marginal distribution of $X_i, 1 \leq i \leq n$. We consider the case for $n = 2$.

From (2.53) it follows that

$$\lim_{x_2 \rightarrow \infty} F_{\mathbf{X}}(x_1, x_2) = F_{X_1}(x_1) \quad (2.54)$$

and

$$\lim_{x_1 \rightarrow \infty} F_{\mathbf{X}}(x_1, x_2) = F_{X_2}(x_2). \quad (2.55)$$

By independence,

$$\lim_{x_2 \rightarrow \infty} F_{\mathbf{X}}(x_1, x_2) = G_{X_1}(x_1) \quad (2.56)$$

and

$$\lim_{x_1 \rightarrow \infty} F_{\mathbf{X}}(x_1, x_2) = G_{X_2}(x_2). \quad (2.57)$$

It then follows from (2.54)-(2.57) that

$$F_{X_i}(x_i) = G_{X_i}(x_i), \quad i = 1, 2. \quad (2.58)$$

Using this in (2.51) it follows that $X_i, 1 \leq i \leq n$ are independent if and only if

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i); \quad (2.59)$$

that is $X_i, 1 \leq i \leq n$, are independent if and only if their joint distribution factors as a product of their marginal distributions. Hence, if we know that $X_i, 1 \leq i \leq n$, are independent random variables, then one can reconstruct their joint distribution by multiplying together their marginal distributions.

When all the random variables $X_i, 1 \leq i \leq n$, have the same distribution, we say that they are *identically distributed*. In addition, if they are independent, they are usually referred to as *independent and identically distributed* (i.i.d.) random variables. For statistical purposes we shall say they are a *random sample of size n* of a random variable X . For computing purposes, it is generally more convenient to work with the joint densities. In this case an equivalent definition of independence (for both continuous and discrete random variables) is that $X_i, 1 \leq i \leq n$, are independent if and only if their joint density factors as a product of all their marginal densities i.e.,

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i). \quad (2.60)$$

Example 2.10 (Joint distributions and independence) Let X_1 and X_2 have the joint pdf

$$f(x_1, x_2) = \begin{cases} e^{-x_1-x_2}, & x_1 > 0, x_2 > 0 \\ 0, & \text{elsewhere.} \end{cases} \quad (2.61)$$

The marginal distribution for X_1 is the integral of this on x_2 :

$$f_{X_1}(x_1) = \int_0^\infty e^{-x_1-x_2} dx_2 = e^{-x_1}, \text{ for } x_1 > 0. \quad (2.62)$$

Moreover, the marginal density of x_2 is e^{-x_2} , for $x_2 > 0$, so that

$$f(x_1, x_2) = f_1(x_1) f_2(x_2),$$

which implies that X_1 and X_2 are independent.

2.7 Expectation

One of the most important operations in probability and statistics is that of finding the average value of a random variable, usually called its *expectation*. The expected value of X is denoted by $E(X)$ and the formulas for its calculation are given by

$$E(X) = \begin{cases} \sum_{x \in \mathbb{R}(X)} x f_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x), & \text{if } X \text{ is continuous.} \end{cases} \quad (2.63)$$

The relation in (2.63) reflects an important rule. That is, most expectation formulas for discrete random variables involve summations while those for continuous random

variables can be obtained by replacing summations with “integrations.” The general rules for calculating remain the same in both cases.

If $g(X)$ is a function of X , then

$$E[g(X)] = \begin{cases} \sum_{x \in \mathbb{R}(X)} g(x) f_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x), & \text{if } X \text{ is continuous.} \end{cases} \quad (2.64)$$

2.7.1 Moments

If a is a real number, then

$$E[(X - a)^n] \quad (2.65)$$

is called the n -th moment of X about a . When $a = 0$, these are called the *moments* of X and if $a = E(X)$, they are usually called the *central moments* of X . In particular, when $a = E(X) \equiv \mu_X$ and $n = 2$,

$$E[(X - \mu_X)^2] \quad (2.66)$$

is called the *variance* of X and is denoted either by $Var(X)$, $\sigma^2(X)$, σ_X^2 or just σ^2 . The square root of σ^2 , σ is called the *standard deviation* of X and is the most commonly used measure of dispersion of X . That is, it measures how much on average, X differs from its *mean value* $E(X)$.

There are a number of important properties of $E(X)$ and $Var(X)$ that we will use repeatedly throughout the text and are stated below for the convenience of the reader.

(1) If a and b are real numbers, then

$$E(aX + b) = aE(X) + b. \quad (2.67)$$

(2) If $X_i, i = 1, 2, \dots, n$ are random variables each having an expectation, then

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i). \quad (2.68)$$

(Properties (1)-(2) are often referred to as the *linearity properties* of $E(X)$.)

(3) If $X_i, i = 1, 2, \dots, n$, are independent random variables, then

$$E(X_1 X_2 \cdots X_n) = E(X_1) E(X_2) \cdots E(X_n). \quad (2.69)$$

For two random variables X and Y a measure of the relation between X and Y is the *covariance*, $Cov(X, Y)$, defined by

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]. \quad (2.70)$$

We note that if X and Y are independent, then $Cov(X, Y) = 0$. More generally, if $Cov(X, Y) = 0$, we say that X and Y are *uncorrelated*. It is important to observe that while independent random variables are uncorrelated, that in general random variables

can be uncorrelated but not independent. In fact, Y can even be a function of X and still be uncorrelated with X . This points out that the notion statistical and functional dependence can be quite different.

Example 2.11 Let X_1 and X_2 have the joint density given in the following Table:

$X_1 \backslash X_2$	-2	-1	1	2	$f_{X_2}(x_2)$
1	0	1/4	1/4	0	1/2
4	1/4	0	0	1/4	1/2
$f_{X_1}(x_1)$	1/4	1/4	1/4	1/4	1

From the marginal densities, $E(X_1) = 5/2$, $E(X_2) = 0$ and $E(X_1X_2) = 0$, which shows that the covariance is zero. However, it is easy to verify that X_1 and X_2 are not independent. In fact, $X_1 = X_2^2$. That is, the value of X_1 completely determines the value of X_2 .

A normalized version of $Cov(X, Y)$ is the *correlation coefficient*;

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}. \quad (2.71)$$

An important property of ρ is that

$$-1 \leq \rho \leq 1 \quad (2.72)$$

with $\rho = \pm 1$ if and only if X and Y are linearly related. Hence, values of ρ close to ± 1 suggests an approximate linear relationship between X and Y , while values of ρ close to zero suggest little or no linear relation between X and Y . This property, as we shall see, is fundamental to interpreting much of the results of regression analysis. In particular, if X and Y are jointly normally distributed, then $\rho = 0$ if and only if X and Y are independent.

Properties of $Var(X)$

From its definition $\sigma^2(X) \geq 0$ and it can be shown that $\sigma^2(X) = 0$ if and only if $P\{X = \mu_X\} = 1$. Furthermore,

$$(1) \quad Var(aX) = a^2 Var(X);$$

$$(2) \quad Var(X + b) = Var(X);$$

$$(3) \quad Var\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 Var(X) + 2 \sum_{i < j} a_i a_j Cov(X_i, X_j).$$

If $X_i, i = 1, 2, \dots, n$, are pairwise uncorrelated, then (3) simplifies to

$$(4) \quad Var\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 Var(X). \quad (2.73)$$

In particular, if $X_i, 1 \leq i \leq n$, are n independent random variables, then (2.73) holds.

Finally, we note some further properties of the covariance.

$$(1) \operatorname{Cov}(X, X) = \operatorname{Var}(X)$$

(2) If $X_i, 1 \leq i \leq n$ and $Y_j, 1 \leq j \leq m$ are random variables, then

$$\operatorname{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \operatorname{Cov}(X_i, Y_j). \quad (2.74)$$

2.7.2 Moment Generating Function

Although moments can often be calculated by summation or integration, these calculations are frequently done more conveniently by a number of indirect approaches. An important technique in this regard is to use the method of generating functions. Here we discuss one class of generating functions, *moment generating functions* (mgf) which are useful not only for the computation of moments, but for determining properties of distributions as well. They will play a key role in our discussion in Chapters 4 and 5 of statistical properties of least squares estimators of regression models.

Definition 2.2 Let X be a random variable. The moment generating function of X , $M_X(t)$ is defined by

$$M_X(t) = E(e^{tX}) \quad (2.75)$$

provided that sum or integral required in (2.75) exists.

As noted in the definition of $M_X(t)$ not all random variables have a well defined moment generating function and if $M_X(t)$ exists, it will only do so for an interval of values of t about $t = 0$. Since we will be mostly concerned with normal random variables, this technical problem will not arise in this text.

If X is a discrete random variable, then

$$M_X(t) = \sum_{x \in \mathbb{R}(x)} e^{tX} f_X(x), \quad (2.76)$$

while if X has a density, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tX} f_X(x) dx. \quad (2.77)$$

By formally differentiating (2.76)-(2.77) we find that

$$\left. \frac{d^n}{dt^n} M_X(t) \right|_{t=0} = E(X^n) \quad (2.78)$$

so that the n -th central moment can be obtained by differentiating $M_X(t)$ n times at $t = 0$. In this sense, $M_X(t)$ generates the moments of X . Several examples of the usefulness of this idea follow.

Example 2.13 Let X be a binomial random variable, then

$$M_X(t) = (pe^t + q)^n. \quad (2.79)$$

From (2.19) and (2.75)

$$M_X(t) = \sum_{x=0}^n e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^n \binom{n}{x} (pe^t)^x q^{n-x}. \quad (2.80)$$

From the binomial theorem it follows that

$$\sum_{x=0}^n \binom{n}{x} (pe^t)^x q^{n-x} = (pe^t + q)^n \quad (2.81)$$

giving (2.79).

Differentiating (2.79) we can obtain formulas for $E(X)$ and $Var(X)$. In fact,

$$E(X) = \frac{d}{dt} M_X(t) \Big|_{t=0} = npe^t (pe^t + q)^{n-1} \Big|_{t=0} = np \quad (2.82)$$

since $p + q = 1$.

Also

$$\begin{aligned} E(X^2) &= \frac{d^2}{dt^2} M_X(t) \Big|_{t=0} \\ &= npe^t (pe^t + q)^{n-1} \Big|_{t=0} + n(n-1)p^2 e^{2t} (pe^t + q)^{n-2} \Big|_{t=0} \\ &= np + n(n-1)p^2. \end{aligned} \quad (2.83)$$

From (2.66)

$$Var(X) = E(X^2) - [E(X)]^2 \quad (2.84)$$

so that

$$Var(X) = np + n(n-1)p^2 - n^2p^2 = np - np^2 = npq. \quad (2.85)$$

We have it as an exercise to compute $E(X)$ and $Var(X)$ by direct summation.

Example 2.14 A random variable X is said to have a *standard normal distribution* if its density is given by

$$f(x) = (2\pi)^{-1/2} e^{-x^2/2}, \quad -\infty < x < \infty. \quad (2.86)$$

The moment generating function of X is given by

$$M_X(t) = e^{t^2/2}. \quad (2.87)$$

To prove (2.87) we will make use of the well known integral

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = 1 \quad (2.88)$$

which establishes that indeed, $f_X(x)$ is a probability density.

Then,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx-x^2/2} dx \quad (2.89)$$

Now,

$$\begin{aligned} -\frac{x^2}{2} + tx &= -\frac{1}{2}(x^2 - 2tx) = -\frac{1}{2}(x^2 - 2tx + t^2 - t^2) \\ &= -\frac{1}{2}(x - t)^2 + \frac{t^2}{2}. \end{aligned} \quad (2.90)$$

Using (2.90) in (2.89) gives

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{t^2/2} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2} dx. \end{aligned} \quad (2.91)$$

Making the substitution $y = x - t$ and using (2.88) gives (2.87).

Using (2.87) one can easily compute $E(X)$ and $Var(X)$. From (2.78) and (2.88)

$$E(X) = \left. \frac{d}{dt} e^{t^2/2} \right|_{t=0} = \left. t e^{t^2/2} \right|_{t=0} = 0 \quad (2.92)$$

so that

$$\begin{aligned} Var(X) &= E(X^2) = \left. \frac{d^2}{dt^2} e^{t^2/2} \right|_{t=0} \\ &= \left. \frac{d}{dt} t e^{t^2/2} \right|_{t=0} = \left. e^{t^2/2} + t^2 e^{t^2/2} \right|_{t=0} = 1. \end{aligned} \quad (2.93)$$

In addition to facilitating the computation of moments, moment generating functions are useful in studying the distribution of the random variables themselves. This follows from the fact that the moment generating function determines the distribution of X itself.

Specifically, if X and Y are two random variables, and $M_X(t) = M_Y(t)$ for $t \in (-a, a)$, then X and Y are identically distributed. This relation is particularly important for determining the distribution of sums of independent random variables. To see this, suppose X_1 and X_2 are independent random variables, then

$$M_{X_1+X_2}(t) = E[e^{t(X_1+X_2)}] = E[e^{tX_1} e^{tX_2}]. \quad (2.94)$$

By (2.69) and (2.75)

$$E[e^{tX_1} e^{tX_2}] = E(e^{tX_1}) E(e^{tX_2}) = M_{X_1}(t) M_{X_2}(t). \quad (2.95)$$

As an example, suppose that $X_i, i = 1, 2$ are independent Poisson random variables with parameters $\lambda_i, i = 1, 2$. Then $X_1 + X_2$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

We begin by finding the mgf of $X_i, i = 1, 2$. From (2.76) and (2.23)

$$\begin{aligned} M_{X_i}(t) &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda_i} \lambda_i^x}{x!} = e^{-\lambda_i} \sum_{x=0}^{\infty} \frac{(e^t \lambda_i)^x}{x!} \\ &= e^{-\lambda_i} e^{\lambda_i e^t}, \quad i = 1, 2. \end{aligned} \quad (2.96)$$

From (2.94)

$$\begin{aligned} M_{X_1+X_2}(t) &= \left(e^{-\lambda_1} e^{\lambda_1 e^t}\right) \left(e^{-\lambda_2} e^{\lambda_2 e^t}\right) \\ &= e^{-(\lambda_1+\lambda_2)} e^{(\lambda_1+\lambda_2)e^t}. \end{aligned} \quad (2.97)$$

Letting $\lambda_3 = \lambda_1 + \lambda_2$ (2.97) is the mgf of a Poisson random variable Y with parameter $\lambda_3 = \lambda_1 + \lambda_2$. Since Y and $X_1 + X_2$ have the same mgf, they have the same distribution.

Property (2.95) extends easily to the situation where $X_i, 1 \leq i \leq n$, are n independent random variables, and $S_n = \sum_{i=1}^n X_i$. Then

$$M_{S_n}(t) = \prod_{i=1}^n M_{X_i}(t). \quad (2.98)$$

In Section 2.8 we will use (2.98) to establish the important fact that the sum of n independent normal random variables is a normal random variable.

2.8 The Normal and Related Random Variables

In statistical theory and in regression analysis in particular, normal random variables play an important role. In this section we will develop some properties of these random variables and several others, the χ^2 (*chi-square*), T and F random variables which are derived from them.

2.8.1 Normal Random Variables

A random variable X is said to be *normally distributed* with parameters (μ, σ^2) - written as $X \sim N(\mu, \sigma^2)$ - if X has a density of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty < x < \infty. \quad (2.99)$$

It can be shown that normal random variables have many important properties. Among these are:

(1) If $a \neq 0$, and $X \sim N(\mu, \sigma^2)$, then

$$aX + b \sim N(a\mu + b, a^2\sigma^2); \quad (2.100)$$

(2) If $X_i, 1 \leq i \leq n$, are independent random variables and $X_i \sim N(\mu_i, \sigma_i^2), 1 \leq i \leq n$, then

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right); \quad (2.101)$$

(3) (Central Limit Theorem) Although (2.101) is an exact relation when the $X_i \sim N(\mu_i, \sigma_i^2)$ and independent, under very general conditions

$$\sum_{i=1}^n a_i X_i \text{ is approximately } N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right), \quad (2.102)$$

provided that n is sufficiently large. (The X 's do not have to be independent for this to hold.) This is one of the most remarkable theorems in probability theory, usually referred to as the *Central Limit Theorem* (CLT). Full details of proofs and conditions under which the theorem holds can be found in Refs. [63, 89].

(4) From (2.99) we find that

$$Z = (X - \mu)/\sigma \quad (2.103)$$

is $N(0, 1)$. Z is usually referred to as a *standard normal random variable*. Conversely, if $Z \sim N(0, 1)$, then $\sigma Z + \mu \sim N(\mu, \sigma^2)$.

From Table A.1 one can easily see that if $Z \sim N(0, 1)$, then

$$P\{-3\sigma \leq Z \leq 3\sigma\} \simeq 1 \quad (2.104)$$

and from (2.99) that for a $N(\mu, \sigma^2)$ random variable

$$P\{\mu - 3\sigma \leq X \leq \mu + 3\sigma\} \simeq 1. \quad (2.105)$$

Thus, it is highly unlikely that an observation of a normal random variable is more than three standard deviations from its mean. In fact, it is quite unlikely that the values of a normal random variable are more than two standard deviations from μ since

$$P\{\mu - 2\sigma \leq X \leq \mu + 2\sigma\} \simeq 0.95. \quad (2.106)$$

This fact is the basis for many “rules of thumb” in statistical analysis.

(5) The normal distribution is a member of the exponential family.

2.8.2 Chi-Square Random Variables

If $X \sim N(0, 1)$, then the random variable $Y = X^2$ has the density

$$f_Y(y) = \begin{cases} (2\pi y)^{-1/2} e^{-y/2}, & y > 0, \\ 0, & y \leq 0. \end{cases} \quad (2.107)$$

The random variable Y is said to have a χ^2 -distribution with one *degree of freedom* (df). If $Y_i, 1 \leq i \leq n$, are independent χ^2 random variables with one df, then their sum

$$\chi^2(n) = \sum_{i=1}^n Y_i \quad (2.108)$$

has the density

$$f_{\chi^2(n)}(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1}, & x > 0, \\ 0, & x \leq 0. \end{cases} \quad (2.109)$$

where $\Gamma(\alpha)$ is the well known *gamma function* defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \alpha > 0 \quad (2.110)$$

and n in (2.109) is a positive integer again called the *degrees of freedom* (df) of the random variable $\chi^2(n)$. A number of useful properties of χ^2 random variables are summarized below.

- (1) If $\chi^2(n_i)$, $1 \leq i \leq m$, are independent χ^2 random variables with n_i df, then their sum

$$\chi^2(n) = \sum_{i=1}^m \chi^2(n_i), \quad n = \sum_{i=1}^m n_i \quad (2.111)$$

is a χ^2 random variable with $n = \sum_{i=1}^m n_i$ degrees of freedom.

- (2) If X is $\chi^2(n)$ and Y is $\chi^2(m)$ and $X = Y + Z$ with Y and Z independent random variables, then $Z = X - Y$ is $\chi^2(n - m)$ if $n > m$.

- (3) $E[\chi^2(n)] = n$ and $Var[\chi^2(n)] = 2n$.

- (4) If n is large, then

$$[\chi^2(n) - n] / \sqrt{2n} \quad (2.112)$$

is approximately $N(0, 1)$. This follows as a consequence of the Central Limit Theorem and the representation given in (2.102).

- (5) If X_i , $1 \leq i \leq n$, are independent and $N(\mu, \sigma^2)$, then $(\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i)$

$$S^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (2.113)$$

is a χ^2 random variable with $n - 1$ degrees of freedom. This is an important result in classical statistics because it shows that the *sample variance*

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (2.114)$$

of n independent $N(\mu, \sigma^2)$ random variables has the distribution of a $\sigma^2 \chi^2(n - 1) / (n - 1)$ random variable. In particular, it follows from (2.114) that $E(s^2) = \sigma^2$. Hence, in statistical terms (see Section 2.10) s^2 is an *unbiased estimator* of the population variance. We will show in Chapters 3 and 5 that this result has an important generalization in estimating the error in regression analysis.

2.8.3 t and F -Distributions

Two other important classes of random variables closely related to the normal are the T and F random variables.

t -Distribution

A T random variable has the density

$$t(x; n) = \frac{\Gamma[(n + 1)/2]}{\Gamma(n/2) \sqrt{n\pi}} \frac{1}{(1 + x^2/n)^{(n+1)/2}}, \quad -\infty < x < \infty \quad (2.115)$$

where n is a positive integer called the *degrees of freedom*. In classical statistical theory T random variables occur naturally in the estimation theory associated with the mean

μ of a normal random variable. As we shall see, they play a similar role in regression analysis.

An examination of Table A.2 suggests that as $n \rightarrow \infty$ the t density is approximately $N(0, 1)$. In fact, with some tedious algebra it can be shown that

$$\lim_{n \rightarrow \infty} t(x; n) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \quad -\infty < x < \infty. \quad (2.116)$$

We will find this result useful in regression analysis because it can be used to give an “eye-ball test” for the significance of regression coefficients when the number of observations is large.

T random variables have their origin in classical statistics. It is well known that if $X_i, 1 \leq i \leq n$, is a random sample of size n of a $N(\mu, \sigma^2)$ random variable, then

$$T = \frac{(\bar{X}_n - \mu) / (\sigma / \sqrt{n})}{\sqrt{s^2 / \sigma^2}} \quad (2.117)$$

has a t density with $n - 1$ degrees of freedom. More generally, if $Z \sim N(0, 1)$ and $\chi^2(m)$ is a chi-square random variable with m df and independent of Z , then

$$T = \frac{Z}{\sqrt{\chi^2(m) / m}} \quad (2.118)$$

has a (Student's) t -distribution with m degrees of freedom. Again, this result, generalizing (2.117) will play a significant role in regression analysis.

F-Distribution

If $\chi^2(n)$ and $\chi^2(m)$ are independent chi-square random variables the random variable

$$F = \frac{\chi^2(n) / n}{\chi^2(m) / m} \quad (2.119)$$

has the density

$$f(x; n, m) = \begin{cases} \frac{(n/m)^{n/2} x^{n/2-1}}{B(n/2, m/2)} [1 + (n/m)x]^{(n+m)/2}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad (2.120)$$

where

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \quad (2.121)$$

is the *beta function*. Such random variables are called “ F ” random variables. (Note: A chi-square random variable divided by its degrees of freedom (df) is often referred to as a *mean square*. Hence, an F random variable is the ratio of two mean squares.) Again the integers (n, m) are called the *numerator* and *denominator* df respectively. In the classical normal statistical theory F random variables occur when one needs to test for the equality of variances. A similar role occurs in regression analysis.

When m is large, then F is approximately $\chi^2(n)$. This can be helpful in finding approximate values for critical points of the F density when the denominator df m is large. This situation is of frequent occurrence in regression analysis because m is a sample size which generally is considerably larger than n .

2.8.4 Lognormal Random Variables

Historically in statistics one often chooses normal random variables to describe data that have a reasonably symmetric density. This choice is frequently justified (without analysis) on the basis of the Central Limit Theorem. However, increasingly one is faced with long-tailed asymmetric data such as stock prices, income levels and housing prices. Frequently, it is argued that by taking the logarithm of the data that the transformed data will be approximately normally distributed. This argument suggests that one consider a random variable Y such that

$$X = \log Y \quad \text{or} \quad Y = e^X \quad (2.122)$$

where X has a $N(\mu, \sigma^2)$ distribution. From (2.122) it is easily shown that Y has a density given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma y} \exp\left[-\frac{1}{2\sigma^2}(\log y - \mu)^2\right], & y > 0, \\ 0, & y \leq 0. \end{cases} \quad (2.123)$$

In this case Y is called a *lognormal random variable*. Although the density is similar in functional form to the normal density, the reader should be cautioned that $\mu \neq E(Y)$ and $\sigma^2 \neq \text{Var}(Y)$. In fact, it can be shown that

$$E(Y) = \exp(\mu + \sigma^2/2) \quad (2.124)$$

and

$$\text{Var}(Y) = \exp(2\mu + \sigma^2) [\exp(\sigma^2) - 1]. \quad (2.125)$$

Note from (2.124) that $E(Y) > e^\mu$ which is the *median* of Y ; a result which reflects the asymmetry of the distribution of Y .

In regression analysis logarithmic transformations of the data are frequently taken (Sec. 3.10) in the hope that the resulting data is normally distributed. This leads to random variables Y that are lognormally distributed. This type of assumption is often used when trying to model data that changes by some multiplicative process over time, such as population, prices, income etc.

2.9 Statistical Estimation

As we shall see, regression analysis is largely concerned with the problem of parameter estimation and the resulting hypothesis testing associated with these estimates. In the current statistical literature, there are a large number of methods for doing this. In this section we will briefly review some of the more common techniques of parameter estimation with particular emphasis on *maximum likelihood estimation* (MLE), a procedure we will find most useful in this text.

2.9.1 The Method of Moments

Suppose that X is a random variable whose density (discrete or continuous) $f_X(x)$ depends on m unknown parameters $(\theta_1, \theta_2, \dots, \theta_m)$. We will denote this dependence by writing the density in the form

$$f_X(x; \theta_1, \theta_2, \dots, \theta_m), \quad (2.126)$$

where the parameter vector $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ lies in some subset $\Theta \subseteq \mathbb{R}^m$ called the *parameter space*. The method of moments uses the sample moments to estimate θ in the following way. Let $X_i, 1 \leq i \leq n$ be a random sample of size n of X . The k -th sample moment of $X_i, 1 \leq i \leq n$ is defined by

$$\overline{X}^{(k)} = \frac{1}{n} \sum_{i=1}^n (X_i)^k. \quad (2.127)$$

If $x_i, 1 \leq i \leq n$ are the actual observed values $X_i, 1 \leq i \leq n$, then

$$\overline{x}^{(k)} = \frac{1}{n} \sum_{i=1}^n (x_i)^k \quad (2.128)$$

is the observed value of $\overline{X}^{(k)}$. To estimate $(\theta_1, \theta_2, \dots, \theta_m)$ we set up the following equations:

$$\overline{x}^{(k_j)} = E(X^{k_j}), \quad 1 \leq j \leq m \quad (2.129)$$

and then solve (2.129) for the m values of the parameters $\theta_i, 1 \leq i \leq m$. The resulting *moment estimators* are denoted by $\hat{\theta}_i, 1 \leq i \leq m$. As is customary in statistics, we will generally not distinguish between the estimators, which are random variables, and their observed values, the estimates, which are real numbers.

Most often, one uses the first m sample moments $\overline{X}^{(k)}, 1 \leq k \leq m$. We illustrate these ideas with a number of examples.

Example 2.15 Suppose that X is a Bernoulli random variable whose density is given by $f_X(0; \theta) = P\{X = 0\} = 1 - \theta$, $f_X(1; \theta) = P\{X = 1\} = \theta$, and $f_X(x; \theta) = 0$, otherwise. As is easily shown, $E(X) = \theta$ and $\overline{X}^{(1)} = \sum_{i=1}^n X_i/n$. Hence, using (2.129) a moment estimator for θ is given by solving

$$\frac{1}{n} \sum_{i=1}^n x_i = E(X) = \theta \quad (2.130)$$

where x_i 's are the observed values of $X_i, 1 \leq i \leq n$. The resulting estimator $\hat{\theta}$ for θ is given by

$$\hat{\theta} = \overline{X}_n, \quad (2.131)$$

which is the *sample mean*.

Example 2.16 Let X be a $N(\mu, \sigma^2)$ random variable (here $\theta_1 = \mu, \theta_2 = \sigma^2$) whose density is given by (2.99). To find moment estimators for (μ, σ^2) we use the fact that $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$. Hence, using (2.127) with the first two moments

$$\mu = \overline{x}_n \quad (2.132)$$

and

$$\overline{x}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n x_i^2 = \sigma^2 + \mu^2. \quad (2.133)$$

Solving (2.133) gives the estimators

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma}^2 = \bar{X}_n^2 - \left[\frac{1}{n} \sum_{i=1}^n X_i^2 \right]. \quad (2.134)$$

Some further algebra shows that

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (2.135)$$

As we shall see, one generally uses $n - 1$ rather than n in (2.135) to define the *sample variance*

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (2.136)$$

which is an *unbiased estimator* for σ^2 .

Example 2.17 In many cases the equations (2.129) cannot be solved explicitly and numerical techniques must be used to obtain explicit values of the estimates. We illustrate this with some data taken from [16]. The problem concerns the distribution of group sizes in various social settings. If one assumes that groups such as pedestrians, groups at parties, etc., form at random, then it is reasonable to assume that the size of such groups is described by a Poisson random variable. However, since a group must contain at least one member, then $P\{X = 0\} = 0$, which is impossible if X is Poisson. Thus we consider the conditional distribution of X conditioned on $X \geq 1$. This distribution is given by (see Exercise 2.14)

$$f(x; \theta) = \frac{e^{-\theta} \theta^x}{x! (1 - e^{-\theta})}, \quad x = 1, 2, \dots \quad (2.137)$$

where $E(X) = \theta / (1 - e^{-\theta})$, $\theta > 0$. Using (2.128) for $k = 1$ we obtain an estimate $\hat{\theta}$ for θ by solving the equation

$$\hat{\theta} / (1 - e^{-\hat{\theta}}) = \bar{x}_n \quad (2.138)$$

where \bar{x}_n denotes the average size of n observed groups.

Using elementary calculus it is easily shown that (2.138) has a unique solution for $\bar{x}_n \geq 1$ and this condition will always be satisfied since $\mathbb{R}(X) \geq 1$. For convenience, let $\bar{x}_n = \mu$ so that (2.138) can be written as

$$\hat{\theta} = \mu (1 - e^{-\hat{\theta}}) \quad (2.139)$$

To solve (2.139) we use a common numerical method, *iteration*. For this we make an initial guess $\hat{\theta}_0$ and successively define the sequence $\hat{\theta}_n$ by

$$\hat{\theta}_{n+1} = \mu (1 - e^{-\hat{\theta}_n}), \quad n \geq 0. \quad (2.140)$$

In Table 2.1 we show the convergence behavior of this method. As one can see, the method converges quite rapidly using the starting values of $\hat{\theta}_0 = \mu$.

Table 2.1 Convergence of $\hat{\theta}_{n+1} = \mu (1 - e^{-\hat{\theta}_n})$

$\hat{\theta}_n \setminus \mu$	1.5	2.0	5.0	7.0	9.0
$\hat{\theta}_0$	1.5000	2.0000	5.0000	7.0000	9.0000
$\hat{\theta}_1$	1.1653	1.7293	4.9636	6.9936	8.9989
$\hat{\theta}_2$	1.0323	1.6452	4.9651	6.9936	8.9989
$\hat{\theta}_3$	0.9657	1.6140	4.9652	6.9936	-
$\hat{\theta}_4$	0.9289	1.6018	-	-	-
$\hat{\theta}_5$	0.9075	1.5969	-	-	-
$\hat{\theta}_6$	0.8947	1.5950	-	-	-
$\hat{\theta}_7$	0.8869	1.5942	-	-	-
$\hat{\theta}_8$	0.8821	1.5938	-	-	-
$\hat{\theta}_9$	0.8790	1.5937	-	-	-
$\hat{\theta}_{10}$	0.8773	1.5936	-	-	-
$\hat{\theta}_{11}$	0.8761	1.5936	-	-	-
$\hat{\theta}_{12}$	0.8750	-	-	-	-
$\hat{\theta}_{13}$	0.8750	-	-	-	-
$\hat{\theta}_{14}$	0.8745	-	-	-	-
$\hat{\theta}_{15}$	0.8744	-	-	-	-

Note: “-” indicates convergence.

Using the estimate $\hat{\theta}$ from (2.140) we can then estimate the density of X by

$$\hat{f}(x; \hat{\theta}) = e^{-\hat{\theta}} \hat{\theta}^x / x! (1 - e^{-\hat{\theta}}). \quad (2.141)$$

In Table 2.2 we show data on shopping group sizes and the predicted values given by (2.141) using $\hat{\theta}_0 = 1.51$ and $\hat{\theta} = 0.889$. As one can see, the fit is quite good. This can be supported further by using the *chi-square test* [89, 63].

Table 2.2 Estimated Frequencies of Shopping Group Sizes

Group size x	Observed frequency of group of size x	Estimated frequency of group of size x
1	316	316
2	141	141
3	44	42
4	5	9
5	4	2

2.9.2 Maximum Likelihood Estimation

In *maximum likelihood estimation* we assume that a sample $X_i, 1 \leq i \leq n$, of a random variable X have a joint density (discrete or continuous) given by

$$L = f(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_m) \quad (2.142)$$

where again $(\theta_1, \theta_2, \dots, \theta_m)$ are the unknown parameters whose estimates are given by choosing $\hat{\theta}_i, 1 \leq i \leq m$ to maximize L for each observed sample sequence (x_1, x_2, \dots, x_n) . In effect, we estimate the parameters in such a way that makes the observed sample the most likely to occur.

When $X_i, 1 \leq i \leq n$ is a random sample of X , then (2.142) can be written as a product

$$L = \prod_{i=1}^n f_X(x_i; \theta_1, \theta_2, \dots, \theta_m). \quad (2.143)$$

In this case it is usually easier to maximize $\log L$ (the *log likelihood function*) rather than the *likelihood function* L itself. If L is sufficiently smooth, then this may be done by solving the simultaneous equations

$$\frac{\partial L}{\partial \theta_i} = 0, \quad 1 \leq i \leq m. \quad (2.144)$$

As for the method of moments (2.144) usually must be solved numerically and there may be more than one solution. The resulting estimators are called *maximum likelihood estimators* (MLE) and have many desirable properties. Hence, they are perhaps the most frequently used estimators in statistical analysis. As we show in Chapters 3 and 5, much of regression analysis is based on this estimation technique. Several examples illustrate this approach.

Example 2.18 Suppose X is a Bernoulli random variable with density given by (2.18). To find the MLE of θ based on a random sample of size n we first observe that $f_X(x)$ can be written as

$$f_X(x; \theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1, \quad 0 \leq \theta \leq 1. \quad (2.145)$$

From (2.143) L can be written as

$$L = \prod_{i=1}^n f_X(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}. \quad (2.146)$$

If $\sum_{i=1}^n x_i = 0$, then $L = (1 - \theta)^n$ which is maximized for $\theta = 0$, while if $\sum_{i=1}^n x_i = n$, $L = \theta^n$ which is maximized for $\theta = 1$. If $0 < \sum_{i=1}^n x_i < n$, then (2.147) can be maximized by solving

$$\frac{\partial \log L}{\partial \theta} = 0. \quad (2.147)$$

From (2.146)

$$\log L = \left(\sum_{i=1}^n x_i \right) \log \theta + \left(n - \sum_{i=1}^n x_i \right) \log (1 - \theta) \quad (2.148)$$

so that

$$\frac{\partial \log L}{\partial \theta} = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{\left(n - \sum_{i=1}^n x_i \right)}{1 - \theta}. \quad (2.149)$$

Setting $\partial \log L / \partial \theta = 0$ and solving for θ gives

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n. \quad (2.150)$$

Using the second derivative test one can show that $\hat{\theta}$ in fact maximizes L . Hence, in all cases $\hat{\theta} = \bar{x}_n$, the sample mean, which is the same as the moment estimator.

Example 2.19 Let X be a $N(\mu, \sigma^2)$ random variable, we find the MLE of (μ, σ^2) in the following way. If $X_i, 1 \leq i \leq n$ is a random sample of X , then using (2.99)

$$\begin{aligned} L &= \prod_{i=1}^n \left\{ \left(\sqrt{2\pi}\sigma \right)^{-1} \exp \left[-(x_i - \mu)^2 / 2\sigma^2 \right] \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \end{aligned} \quad (2.151)$$

so that

$$\log L = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (2.152)$$

As before, we can maximize $\log L$ using calculus by setting $\partial \log L / \partial \mu = \partial \log L / \partial \sigma = 0$. Doing this gives

$$\sum_{i=1}^n (x_i - \mu) = 0 \quad (2.153)$$

and

$$\frac{n}{\sigma} - \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3} = 0. \quad (2.154)$$

From (2.153) $\hat{\mu} = \bar{x}_n$ and using this in (2.154) gives

$$\hat{\sigma} = \left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right]^{1/2}. \quad (2.155)$$

Notice that in (2.155) that the MLEs agree with the corresponding moment estimators. This is not always the case. For example, suppose X is a uniform random variable with a density

$$f_X(x) = 1/\theta, \quad 0 \leq x \leq 1 \quad (2.156)$$

then the method of moments estimator (MME) of θ is given by

$$\hat{\theta}_{MME} = 2\bar{X}_n \quad (2.157)$$

while the MLE is given by

$$\hat{\theta}_{MLE} = \max(X_1, X_2, \dots, X_n). \quad (2.158)$$

Generally, one prefers $\hat{\theta}_{MLE}$ to $\hat{\theta}_{MME}$ since it has smaller variance - hence it is more efficient (see Section 2.10 for the definition). The efficiency of MLEs is one of the properties that makes them so desirable in practice.

2.9.3 Least Squares Estimation

In least squares estimation we assume that $X_i, 1 \leq i \leq n$, are n random variables with

$$E(X_i) = f_i(\theta_1, \theta_2, \dots, \theta_m). \quad (2.159)$$

If $x_i, 1 \leq i \leq n$ are the observed values of X_i , then the least squares estimates of $(\theta_1, \theta_2, \dots, \theta_m)$ are obtained by minimizing

$$Q = \sum_{i=1}^n [x_i - f_i(\theta_1, \theta_2, \dots, \theta_m)]^2 \quad (2.160)$$

with respect to $(\theta_1, \theta_2, \dots, \theta_m)$. If the $f_i, 1 \leq i \leq n$ are sufficiently smooth, then this may be done by calculus by solving

$$\partial Q / \partial \theta_i = 0, \quad 1 \leq i \leq m. \quad (2.161)$$

As for maximum likelihood estimation, this usually must be done numerically. As we shall see, in classical regression analysis MLEs and least squares estimators (LSEs) are often the same.

2.9.4 Bayesian Estimation

In the method of moments and maximum likelihood estimation the unknown parameter(s) $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ is assumed to be a fixed number subject only to the condition that $\theta \in \Theta$, the prescribed parameter space. In these forms of estimation all possible values of θ are treated on the same footing before estimates are made. However, in many situations it is reasonable to assume on the basis of past experience that some values of θ are more likely to occur than others. For example, in tossing a coin, which is described by a Bernoulli random variable with $\theta = P\{X = 1\}$ ($\theta = P\{\text{Heads occurs}\}$) that values of θ near one-half are more likely to occur than those near zero or one. In this situation we can think of θ as a random variable and then $f_X(x_i; \theta)$ can be interpreted as the conditional density of X given θ . If θ has a density, then

$$P\{X = x\} = \int_{\Theta} f_X(x; \theta) f(\theta) d\theta \quad (2.162)$$

for a discrete random variable and

$$P\{x \leq X \leq x + \Delta x\} \simeq \Delta x \int_{\Theta} f_X(x; \theta) f(\theta) d\theta \quad (2.163)$$

for small Δx if X is continuous.

In Bayesian estimation observations are used to modify the *prior distribution* of θ via *Bayes rule*. Characteristics of this *posterior distribution* are then used to estimate θ . As the theory can become quite complex, we will only illustrate the rudiments of the approach.

If $X_i, 1 \leq i \leq n$, is a random sample of X , Bayes rule yields the *posterior density* (discrete or continuous)

$$f_{\theta|x}(\theta|x_1, x_2, \dots, x_n) = \frac{\prod_{i=1}^n f_X(x_i; \theta)}{\int_{\Theta} [\prod_{i=1}^n f_X(x_i; \theta)] f(\theta) d\theta}. \quad (2.164)$$

To estimate θ one now chooses some characteristic of $f_{\theta|\mathbf{x}}(\theta|x_1, x_2, \dots, x_n)$ such as its mean, mode (most likely value), median, etc. Choosing the mode, for example (if it exists) yields an estimate of θ as the “most likely” value of θ given the observations (MLE). Justification for using the mean or median can be given in terms of statistical decision theory. The mean of the posterior distribution is the most commonly used estimate.

Example 2.20 Again we consider the problem of estimating the parameter of a Bernoulli distribution. Assume now that θ is a random variable and $\mathbb{R}(\theta) = [0, 1]$. To apply (2.164) requires the choice of a prior density $f_{\theta}(\theta)$. Suppose we are completely ignorant of the possible values of θ , i.e., all values of θ are assumed equally likely. In this case we choose θ to be uniform on $[0, 1]$. The posterior distribution is then

$$\begin{aligned} f_{\theta|\mathbf{x}}(\theta|x_1, x_2, \dots, x_n) &= \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}}{\int_0^1 \left[\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \right] d\theta} \\ &= \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}}{B(\sum_{i=1}^n x_i + 1, n + 1 - \sum_{i=1}^n x_i)} \end{aligned} \quad (2.165)$$

where the denominator in (2.165) is a beta function. Since the denominator is independent of θ choosing $\hat{\theta}$ as the mode of the posterior distribution yields the maximum likelihood estimator \bar{X}_n .

If we use $\hat{\theta} = E(\theta|\mathbf{X})$, the *posterior mean*, then using properties of the beta function gives

$$\hat{\theta} = E(\theta|\mathbf{X}) = \frac{\sum_{i=1}^n X_i + 1}{n + 2}, \quad (2.166)$$

which differs from \bar{x}_n . For example, if $\sum_{i=1}^n x_i = 0$ then $\hat{\theta} = 1/(n+2)$ and not zero. For small values of n the Bayes estimator tends to pull in extreme values of \bar{X}_n towards the center of $[0, 1]$. This is sensible since for small sample sizes there is a reasonable probability of getting strings of all zeros or ones. For large values of n the Bayes estimator $\hat{\theta}$ and \bar{X}_n differ by little.

If complete ignorance does not prevail, then we will feel that some values of θ are more likely than others and we are faced with translating our subjective feeling into quantitative statements about a prior distribution for θ . A typical choice is to assume that θ has a beta density (this makes calculations easy to do) so that

$$f_{\theta|\mathbf{x}}(\theta|\mathbf{x}) = \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)}. \quad (2.167)$$

If one chooses $\hat{\theta}$ as the mode of (2.167), then if $\alpha \geq 1$, $\beta \geq 1$,

$$\hat{\theta} = \frac{\sum_{i=1}^n X_i + \alpha - 1}{\alpha + \beta + n - 2}, \quad (2.168)$$

whereas using the posterior mean yields

$$\hat{\theta} = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}. \quad (2.169)$$

For large enough n all of these estimators will differ little from \bar{X}_n .

Example 2.21 Let X be a $N(\theta, \sigma^2)$ random variable, and suppose that the prior distribution of θ is $N(\eta, \tau^2)$. Assuming that σ^2, η , and τ^2 are known, then the posterior distribution of θ is also normal, with mean and variance given by (after some algebra)

$$E(\theta|x) = \left(\frac{\tau^2}{\sigma^2 + \tau^2} \right) x + \left(\frac{\sigma^2}{\sigma^2 + \tau^2} \right) \eta, \quad (2.170)$$

$$Var(\theta|x) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \quad (2.171)$$

Thus the posterior mean, which is the Bayes estimator, is a linear combination of the prior mean and the sample means.

It is worth noting that if we allow τ^2 , the prior variance, to tend to infinity, the Bayes estimator tends towards the sample mean. This implies that as the prior information is getting more vague; the Bayes estimator tends to give more weight to the sample information.

2.10 Properties of Estimators

2.10.1 Consistency and Unbiasedness

Having presented several techniques for making parameter estimates and observing that different methods may produce different estimators, we now turn our attention to a discussion of some of the more important criteria which are used to evaluate competing ones.

Since estimators are random variables, we can examine their distributions for clues to their characteristics. Since the purpose of sampling is to determine the true value of θ , it is reasonable to assume that as the sample size n increases the distribution of $\hat{\theta}_n$ clusters more about the true values of θ for all $\theta \in \Theta$. This may be expressed in probability terms as

$$\lim_{n \rightarrow \infty} P \left\{ \left| \hat{\theta}_n - \theta \right| < \varepsilon \right\} = 1 \quad (2.172)$$

for any $\varepsilon > 0$. (This is usually called *convergence in probability*.) If (2.172) holds, we shall call $\hat{\theta}_n$ a *consistent* sequence of estimators for θ .

Since one can rarely compute the exact distribution of $\hat{\theta}_n$, we need a criterion which can be easily used to establish consistency. It can be shown that sufficient conditions for (2.172) to hold are that

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta \text{ and } \lim_{n \rightarrow \infty} Var(\hat{\theta}_n) = 0. \quad (2.173)$$

A proof of this can be found in [40, 89].

One should note that if $E(\hat{\theta}_n) = \mu_n = \theta$ for all n ($\hat{\theta}_n$ is then *unbiased*) then it is sufficient that $Var(\hat{\theta}_n) \rightarrow 0$ in order that the sequence $\hat{\theta}_n$ be consistent.

In general, the quantity $\theta - \mu_n$ is called the *bias* of the estimator and

$$E \left[\left(\hat{\theta}_n - \theta \right)^2 \right] = Var(\hat{\theta}_n) + (\theta - \mu_n)^2 \quad (2.174)$$

is called the *mean square error* (MSE) of $\hat{\theta}_n$. Thus an equivalent way of stating (2.174) is to say that $\{\hat{\theta}_n\}$ is consistent if the mean square error converges to zero. Since the rate of convergence of $\hat{\theta}_n$ to θ depends on the mean square error, it is appropriate to look for estimates which minimize it. This can be done, for example, by choosing $\hat{\theta}_n$ to be unbiased and then picking $\text{Var}(\hat{\theta}_n)$ as small as possible. Such an estimator, if it exists, is called a *minimum variance unbiased estimator* (MVUE) of θ . If such an estimator $\hat{\theta}_n$ exists, then we can define the efficiency of any other unbiased estimator $\hat{\psi}_n$ by

$$\text{Efficiency} = \text{Var}(\hat{\theta}_n) / \text{Var}(\hat{\psi}_n). \quad (2.175)$$

More generally, we can say that for any two unbiased estimators $\hat{\theta}$ and $\hat{\psi}$ that $\hat{\theta}$ is more *efficient* than $\hat{\psi}$ if $\text{Var}(\hat{\theta}) < \text{Var}(\hat{\psi})$.

These observations lead to an additional principle of estimation, that of minimizing the MSE, or the variance if we want the estimate to be unbiased. In general, it is quite difficult to find minimum variance estimators and the general theory is beyond the scope of this text. (However, we will consider some special cases in Chapters 3 and 5.)

Example 2.22 Let X be a random variable with density $f(x; \theta)$ such that $E(X) = \theta$. If $\text{Var}(X) < \infty$, then $\sum_{i=1}^n X_i/n$ is a consistent sequence of estimators for θ if $X_i, 1 \leq i \leq n$, is a random sample of X .

Using the linearity of expectation,

$$E(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{n\theta}{n} = \theta \quad (2.176)$$

so \bar{X}_n is unbiased. From (2.73)

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} [n \text{Var}(X)] \\ &= \frac{1}{n} \text{Var}(X) \rightarrow 0, \quad n \rightarrow \infty. \end{aligned} \quad (2.177)$$

As particular cases, \bar{X}_n is a consistent estimator of the mean μ of a $N(\mu, \sigma^2)$ random variable and for the parameter θ of a Bernoulli random variable.

Example 2.23 Show that the sample variance $\hat{\sigma}_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 / n$ is a consistent sequence of estimators for the variance of a normal random variable.

From (2.113) we know that $\sum_{i=1}^n (X_i - \bar{X}_n)^2 / \sigma^2$ is a $\chi^2(n-1)$ random variable so that

$$E(\hat{\sigma}_n^2) = (n-1)\sigma^2/n \quad (2.178)$$

and

$$\text{Var}(\hat{\sigma}_n^2) = 2\sigma^4(n-1)/n^2. \quad (2.179)$$

From (2.178) and (2.179)

$$\lim_{n \rightarrow \infty} E(\hat{\sigma}_n^2) = \sigma^2 \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\sigma}_n^2) = 0 \quad (2.180)$$

and (2.180) shows that $\{\hat{\sigma}_n^2\}$ is a consistent sequence of estimators for σ^2 .

2.10.2 Sufficiency

Although consistency is a basic property which good estimators should have, it is a large sample property, and it is also important to have estimators which exhibit good properties for small values of n . One of these is *unbiasedness*, another one is *sufficiency*.

To illustrate the concept of sufficiency we again consider the problem of estimating θ , the probability of success on a single Bernoulli trial. If $X_i, i = 1, 2, \dots, n$, is a random sample of size n , then we can ask what is the best way to summarize the sample data so that no information about estimating θ is lost. A useful way of looking at this problem is to consider what happens if we want to store this information in a computer. Assume that all numbers are stored in binary so that recording all the outcomes gives rise to the binary number $\epsilon_1\epsilon_2 \cdots \epsilon_n$ where ϵ_i is 0 or 1. Thus storing all the information about the experiment requires n "bits". In Section 2.9 we have observed that \bar{X}_n is in many senses a good estimator for θ and to compute \bar{X}_n we need only store the number of successes $\sum_{i=1}^n X_i$. This requires on the order of $\log_2 n$ bits, rather than n .

For example, suppose we toss a coin 10 times and observe four H 's and six T 's. Then storing $(x_1, x_2, \dots, x_{10})$ requires 10 bits. Now $\sum_{i=1}^{10} x_i = 4$ and $4 = 100$ in binary and so its storage requires only three bits. Even if $\sum_{i=1}^{10} x_i = 10$ then $10 = 1010$ (in binary) and we would need at most four bits. Since $\lim_{n \rightarrow \infty} \log_2 n/n = 0$ the information compression given by $\sum_{i=1}^n X_i$ becomes arbitrarily large. In fact, it turns out that we cannot do any better, in the sense that $\sum_{i=1}^n X_i$ summarizes all the information about the experiment for the purpose of estimating θ . Nothing further about $\{X_i\}$ such as the order of the outcomes needs to be used to estimate θ . In this case we say that $\sum_{i=1}^n X_i$ is sufficient for θ .

Establishing sufficiency of an estimator can be quite difficult. However, there is a famous theorem, the *factorization theorem* which often simplifies matters considerably [63, 89].

Theorem 2.1 (*Factorization Theorem*) Let X_1, X_2, \dots, X_n be a random sample of size n from the density $f(x; \theta)$. Then $\hat{\theta}(X_1, X_2, \dots, X_n)$ is a sufficient statistic for θ if and only if the joint density $\prod_{i=1}^n f_X(x_i; \theta) \equiv f_{\mathbf{X}}(x_1, x_2, \dots, x_n; \theta)$ factors as follows:

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n; \theta) = g\left[\hat{\theta}(x_1, x_2, \dots, x_n); \theta\right] h(x_1, x_2, \dots, x_n) \quad (2.181)$$

where h does not depend on θ .

Proof. See [63, 89]. ■

Since a sufficient statistic summarizes all the sample information for estimating θ , it is often advisable to begin the search for a goal estimator by checking to see if a sufficient statistic exists. Since a monotone function of a sufficient statistic is again sufficient, one can then modify a given sufficient statistics to have other properties such as unbiasedness.

Example 2.24 Let X_1, X_2, \dots, X_n be a random sample of size n from a $N(\theta, 1)$ population. Then $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is sufficient for θ , where $X_i, 1 \leq i \leq n$, is a random sample of X .

To show this we use (2.181). Hence,

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n; \theta) = \frac{1}{(\sqrt{2\pi})^n} \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \right] \quad (2.182)$$

and using $\sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \theta)^2$. We find that the right hand side of (2.182) becomes

$$\frac{1}{(\sqrt{2\pi})^n} \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right] \exp \left[-\frac{n}{2} (\bar{x}_n - \theta)^2 \right]. \quad (2.183)$$

Letting $h(x_1, x_2, \dots, x_n) = \exp \left[-\sum_{i=1}^n (x_i - \bar{x}_n)^2 / 2 \right]$ and $g(x_1, x_2, \dots, x_n; \theta) = (2\pi)^{-n/2} \exp \left[-n(\bar{x}_n - \theta)^2 / 2 \right]$ in (2.183) shows that \bar{X}_n is a sufficient statistic for θ .

The factorization criterion can be extended to the problem of jointly estimating m parameters $(\theta_1, \theta_2, \dots, \theta_m)$.

Definition 2.3 Let X_1, X_2, \dots, X_n be a random sample of size n from the density $f_X(x; \theta_1, \theta_2, \dots, \theta_m)$ that depends on m parameters. Then the statistics $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$ are *jointly sufficient* for $(\theta_1, \theta_2, \dots, \theta_m)$ if and only if the joint density of a random sample of X factors as

$$f_{\mathbf{X}}(x_1, \dots, x_n; \theta_1, \dots, \theta_m) = g\left\{\hat{\theta}_1(\mathbf{x}), \dots, \hat{\theta}_m(\mathbf{x}); \theta\right\} h(x_1, \dots, x_n) \quad (2.184)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Example 2.25 Let X_1, X_2, \dots, X_n be a random sample of size n from a $N(\mu, \sigma^2)$. Then \bar{X}_n and $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are jointly sufficient for estimating (μ, σ^2) . This follows from (2.184) since the joint density is

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2, \dots, x_n; \mu, \sigma^2) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2 \right\} \right]. \end{aligned} \quad (2.185)$$

From this we see that

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n; \mu, \sigma^2) = g \left[\bar{x}_n, \sum_{i=1}^n (x_i - \bar{x}_n)^2; \mu, \sigma^2 \right]. \quad (2.186)$$

Thus (2.186) is satisfied by taking $h(x_1, x_2, \dots, x_n) = 1$.

Since a sufficient statistic summarizes all the data for θ one can expect that “best” estimators for θ should be a function of a sufficient statistic. In fact it can be shown that if $\hat{\theta}$ is an unbiased estimator for θ then $E(\hat{\theta}|\hat{\theta}_s)$ is also an unbiased estimator, where $\hat{\theta}_s$ is sufficient, and has variance no greater than $\text{Var}(\hat{\theta})$. Thus in order to find minimum variance estimators it suffices to look for those which are functions of sufficient statistics. For further results along these lines the reader is referred to Ref. [74].

2.11 Confidence Intervals

2.11.1 Exact Confidence Intervals

Up to this point we have examined methods for finding point estimates of parameters for a given random variable. However, as in non-statistical calculations, when an approximation is made it is important to have some estimate of the error in the calculation. For example, if we wish to estimate some real number x by the number \hat{x} and Δx is the absolute error then we will have

$$\hat{x} - \Delta x \leq x \leq \hat{x} + \Delta x. \quad (2.187)$$

That is, the true value x lies in the interval $[\hat{x} - \Delta x, \hat{x} + \Delta x]$. In statistical estimation if θ is the parameter of a random variable and $\hat{\theta}$ is an estimator, then the interval $[\hat{\theta} - \Delta\theta, \hat{\theta} + \Delta\theta]$ has random end points and so we cannot make error statements with probability one, but only with a given level of confidence $1 - \alpha$. Moreover, we need to make probability statements which do not depend on the unknown parameter θ . Formalizing these considerations leads us to the notion of a *confidence interval* (or interval estimate) for a parameter θ .

Definition 2.4 Let X be a random variable whose distribution depends on a parameter θ . A confidence interval for θ with confidence at least $(1 - \alpha) \times 100\%$ is a pair of statistics $(\hat{\theta}_L, \hat{\theta}_U)$, $\hat{\theta}_L \leq \hat{\theta}_U$, such that for all $\theta \in \Theta$

$$P\{\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\} \geq 1 - \alpha. \quad (2.188)$$

If the inequality in (2.188) can be taken to be an equality, we say that $(\hat{\theta}_L, \hat{\theta}_U)$ is an *exact* $(1 - \alpha) \times 100\%$ confidence interval for θ .

For continuous random variables one can often find exact confidence intervals, whereas for discrete random variables one can usually only satisfy (2.188) with the inequality.

A standard technique for finding confidence intervals is based on the notion of a *pivotal quantity* Q where Q is a function of a random sample (X_1, X_2, \dots, X_n) of X whose distribution does not depend on the unknown parameters in the distribution or density of X . Q itself is not a statistic since it will usually be a function of the unknown parameters. To illustrate these ideas consider the problem of finding a confidence interval for the mean μ of a $N(\mu, \sigma^2)$ random variable and suppose for the time being that σ is known. Then from previous considerations we base our estimation of μ on \bar{X}_n . Now, $Q_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ is $N(0, 1)$ and so can serve as a pivotal quantity. Thus we can determine (a, b) independent of (μ, σ^2) such that

$$P\{a \leq Q_n \leq b\} = 1 - \alpha. \quad (2.189)$$

Manipulating the inequality in (2.189) gives

$$P\{\bar{X}_n - b\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n - a\sigma/\sqrt{n}\} = 1 - \alpha \quad (2.190)$$

so that $(\bar{X}_n - b\sigma/\sqrt{n}, \bar{X}_n - a\sigma/\sqrt{n})$ is a $(1 - \alpha) \times 100\%$ confidence interval for μ . Since a and b are not unique, we have in fact found infinitely many confidence intervals for μ .

To make $\hat{\theta}_U - \hat{\theta}_L$ as “short” as possible a and b should be chosen to minimize $E(\hat{\theta}_U - \hat{\theta}_L)$ and this is easily shown to result in the choice $a = -b$. Thus,

$$P\{-b \leq Q_n \leq b\} = 1 - \alpha, \quad (2.191)$$

and this gives $b = z_{\alpha/2}$ where $P\{Q_n \geq z_{\alpha/2}\} = \alpha/2$ and $z_{\alpha/2}$ is the $[(1 - \alpha/2) \times 100\%]$ -th percentage point of a $N(0, 1)$ random variable. This gives the confidence interval as

$$(\bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}). \quad (2.192)$$

Example 2.26 (Confidence interval for σ^2) Let X_1, X_2, \dots, X_n be a random sample of size n from a $N(\mu, \sigma^2)$. Find a $(1 - \alpha) \times 100\%$ confidence interval for σ^2 if μ is unknown.

By (2.113) we know that

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (2.193)$$

is a $\chi^2(n-1)$ random variable. Since the density of $\chi^2(n-1)$ does not depend on σ^2 , we can use χ^2 as a pivotal quantity. Thus we choose (a, b) so that

$$P\{a \leq \chi^2 \leq b\} = 1 - \alpha. \quad (2.194)$$

One solution to (2.194) (the customary one) is to determine b by $P\{\chi^2(n-1) \geq b\} = \alpha/2$ and a by $P\{\chi^2(n-1) \leq a\} = \alpha/2$. If we let $\chi_{n-1, \alpha/2}^2 = b$ and $\chi_{n-1, 1-\alpha/2}^2 = a$ then

$$P\left\{\chi_{n-1, 1-\alpha/2}^2 \leq \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \leq \chi_{n-1, \alpha/2}^2\right\} = 1 - \alpha. \quad (2.195)$$

After a little manipulation, (2.195) gives

$$\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 / \chi_{n-1, \alpha/2}^2, \sum_{i=1}^n (X_i - \bar{X}_n)^2 / \chi_{n-1, 1-\alpha/2}^2\right) \quad (2.196)$$

as a $(1 - \alpha) \times 100\%$ confidence interval for σ^2 .

We should point out that confidence intervals derived under the assumption of normality are often used even if this condition is not met. Usually this will prove satisfactory if the sample size is large (> 30) or if the distribution of X is not highly skewed. This property is termed *robustness*.

2.11.2 Approximate Confidence Intervals

In all the cases considered so far we have been able to find exact confidence intervals for the parameters at hand. In many cases exact confidence intervals are not easily found either because the underlying random variable is discrete or because no convenient pivotal quantity can be found. However, if the sample size is large we may often find random variables whose distribution or density is approximately independent of the given

parameters and so may be used as approximate pivotal quantities. These may then be used to give approximate confidence intervals for a given parameter.

As a typical example suppose that $\theta = E(X)$ is the expected value of X . If $Var(X) < \infty$ is known and $X_i, i = 1, 2, \dots, n$ is a random sample of size n of X , then it follows from the Central Limit Theorem (CLT) that for large n that $Z_n = \sqrt{n}(\bar{X}_n - \theta)/\sigma$ is approximately $N(0, 1)$ and so may be used as an approximate pivotal quantity for θ . Proceeding as in the preceding section we find that

$$(\bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}) \quad (2.197)$$

is an approximate $(1 - \alpha) \times 100\%$ confidence interval for θ .

If $Var(X) = \sigma^2$ is unknown, then as we mentioned in Section 2.8, Eq. (2.117) may be used to find an approximate confidence interval for μ . Since T random variables are approximately normal for large n (≥ 25), we get the further approximation

$$(\bar{X}_n - z_{\alpha/2}s_n/\sqrt{n}, \bar{X}_n + z_{\alpha/2}s_n/\sqrt{n}) \quad (2.198)$$

as an approximate $(1 - \alpha) \times 100\%$ confidence interval for μ .

Example 2.27 Let X_1, X_2, \dots, X_n be a random sample of size n from the Bernoulli distribution with parameter θ . For large n find an approximate confidence interval for $\theta = P\{X = 1\}$.

Define $\bar{X}_n = \sum_{i=1}^n X_i/n$. From the Central Limit Theorem, $\sqrt{n}(\bar{X}_n - \theta)/\sqrt{\theta(1 - \theta)}$ is approximately $N(0, 1)$. Thus,

$$(\bar{X}_n - z_{\alpha/2}\sqrt{\theta(1 - \theta)/n}, \bar{X}_n + z_{\alpha/2}\sqrt{\theta(1 - \theta)/n}) \quad (2.199)$$

is an approximate $(1 - \alpha) \times 100\%$ confidence interval for θ .

Since the end points in (2.199) depend on θ , which is unknown, (2.199) cannot be used as it stands. Although (2.199) can be manipulated to produce a legitimate confidence interval, it is simpler to replace θ with its unbiased estimate \bar{X}_n (or frequently denoted by $\hat{\theta}$ and called the *sample proportion*). This gives the further useful approximate confidence interval

$$(\bar{X}_n - z_{\alpha/2}\sqrt{\bar{X}_n(1 - \bar{X}_n)/n}, \bar{X}_n + z_{\alpha/2}\sqrt{\bar{X}_n(1 - \bar{X}_n)/n}) \quad (2.200)$$

for θ .

2.12 Hypothesis Testing

In the estimation problem experiments are conducted for the purpose of trying to determine the value of some unknown parameter θ . If we have some preconceived idea of what the parameter should be, we may want to conduct an experiment to either confirm our belief about θ or reject the hypothesized value of θ . For example, suppose we are given a coin which we feel is fair, but to be on the safe side we decide to conduct an experiment to see if this is true. Here we hypothesize that $P(\text{Heads}) = 1/2$ and after

performing the experiment we will make one of two decisions. We can accept the hypothesis that $P(\text{Heads}) = 1/2$ or reject this and accept the alternative possibility that $P(\text{Heads}) \neq 1/2$.

A general version of this type of problem may be given in the following terms. Suppose that X is either a discrete or an continuous random variable with density $f_X(x; \theta)$. Suppose we know θ can take values in one of two possible sets Θ_0 and Θ_1 . An experiment is performed to measure X and on the basis of the results we will make one of two decisions; either $\theta \in \Theta_0$ or $\theta \in \Theta_1$.

The statement that $\theta \in \Theta_0$ will be called the *null hypothesis* which is customarily denoted by H_0 . The statement that $\theta \in \Theta_1$ is called the *alternative hypothesis* and is denoted by H_1 . In customary terms we say we are going to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$.

In order to test the hypothesis suppose we take a random sample of size n of X . If $\mathbb{R}^n(X)$ denotes the sample space, we will decide in favor of H_0 if $\mathbf{x} = (x_1, x_2, \dots, x_n) \in S_0 \subseteq \mathbb{R}^n(X)$ and decide in favor of H_1 if $\mathbf{x} \in S_1 = \overline{S_0}$. The partition (S_0, S_1) of the sample space $\mathbb{R}^n(X)$ is called a *test* for the hypothesis H_0 against H_1 . The theory of hypothesis testing is essentially concerned with devising “good” tests.

Example 2.28 A professor has observed that over the past 10 years his probability class has achieved a mean grade of 76 on the final exam and the Dean has suggested that he make greater use of audio-visual aids in order to have his class achieve higher grades. The professor feels that if he uses such aids the mean grade will not be decreased so he decides to do this next semester to see what happens. Describe the decision that the professor faces as a hypothesis testing problem.

Let us assume that the final grade of the typical student can be described as a $N(\theta, \sigma^2)$ random variable. then the professor wants to decide whether $\theta = 76$ or $\theta > 76$. Here we have $\Theta_0 = \{\theta = 76\}$ and $\Theta_1 = \{\theta > 76\}$. Thus $H_0 : \theta \in \{76\}$ and $H_1 : \theta \in \{\theta > 76\}$.

As a test we choose $S_1 = \{\overline{X}_n \geq k\}$ and $S_0 = \{\overline{X}_n < k\}$ for some k . (Justify this.)

2.12.1 Best Tests

To devise good tests we must consider the nature of the errors that can be made in deciding between H_0 and H_1 . We begin with the special case where θ_0 and θ_1 each consist of one point. This is customarily called the problem of testing a *simple hypothesis* against a *simple alternative*. (If either θ_0 or θ_1 consists of more than one point then we say that the corresponding hypothesis is *composite*.)

Since either H_0 or H_1 can be true there are two types of errors that we can make. We can accept H_0 when H_1 is true or we can accept H_1 when H_0 is true. Since we accept H_0 when $\mathbf{x} \in S_0$ we decide in favor of H_1 when $\mathbf{x} \in S_1 = \overline{S_0}$. Thus the probability of deciding in favor of H_1 when H_0 is true is $P_0(S_1) = P(S_1|H_0)$ where $P_0(S_1)$ is the probability of S_1 using the measure induced on $\mathbb{R}^n(X)$ by $f_X(x; \theta_0)$. If we decide in favor of H_1 then $\mathbf{x} \in S_1$, so the probability of deciding in favor of H_0 when H_1 is true is $P(S_0|H_1)$. S_1 is called the *critical region* of the test (i.e., the region where H_0 is rejected) and $P(S_1|H_0) = \alpha$ is called the *size of Type I error*. $P(S_0|H_1) = \beta$ is called the *size of Type II error*. Since α and β are the probabilities of making errors, a good test should try to make these as small as possible. Ideally we would like to minimize α and β simultaneously. Unfortunately this cannot be done. In the extreme case, if we

take $S_0 = \mathbb{R}^n(X)$ then $\alpha = P\{\overline{S_0} = \phi | H_0\} = 0$ while $\beta = P\{\mathbb{R}^n(X) | H_1\} = 1$. Thus minimizing α maximizes β . Hopefully, we can find some happy medium.

In the most widely used version of the hypothesis testing problem α is fixed and we try to find a test which minimizes β . A test which minimizes β from among all those which have the size of Type I error at most α is called a *best test*. Theorem 2.2, the famous *Neyman-Pearson lemma*, shows how to construct a best test for a simple hypothesis against a simple alternative.

Definition 2.5 If a test has type I error of size α , then the critical region S_1 is said to be of *size* α . α is also called the *significance level* of the test.

Theorem 2.2 (Neyman-Pearson Lemma) If there exists a critical region S_1 of size α and a non-negative constant k such that

$$L = \frac{\prod_{i=1}^n f_{\mathbf{x}}(x_i; \theta_1)}{\prod_{i=1}^n f_{\mathbf{x}}(x_i; \theta_0)} \geq k, \mathbf{x} \in S_1 \quad (2.201)$$

and

$$L < k, \mathbf{x} \in S_0, \quad (2.202)$$

then (S_0, S_1) is a best test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. (An intuitive interpretation of (2.201) and (2.202) is that we reject H_0 if the “probability” of $\mathbf{x} = (x_1, x_2, \dots, x_n)$ occurring under H_1 is much larger than the probability of \mathbf{x} occurring under H_0 .)

Proof. A proof of Theorem 2.2 can be found in [63, 89]. ■

The ratio L in (2.201) is usually called the *likelihood ratio* and the test given by (2.202) is a *likelihood ratio test* (LRT). Before giving several specific applications of Theorem 2.2 let us outline the mechanics of its use.

In general the inequality $L \geq k$ is quite complicated so that we begin to determine the critical region by manipulating $L \geq k$ so that S_1 is given by a more tractable inequality $L' \geq k'$. Since L , or equivalently, L' is a function of \mathbf{x} , it is a random variable, and then k' is chosen so that $P\{L' \geq k' | H_0\} = \alpha$. In order to do this we need to be able to calculate the distribution of L' under H_0 . In special cases this can be done exactly. More often, one usually has to rely on approximations, such as the Central Limit Theorem.

Example 2.29 Let X_1, X_2, \dots, X_n be a random sample of size n from a $N(\theta, 1)$. Find the best test based on a random sample of size n of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$.

Here, $f_{\mathbf{x}}(\mathbf{x}; \theta_i) = \exp \left[-\sum_{j=1}^n (x_j - \theta_i)^2 / 2 \right] / (2\pi)^{n/2}$, $i = 0, 1$. Thus, the likelihood

ratio

$$\begin{aligned}
 L &= \exp \left[\frac{1}{2} \left\{ - \sum_{j=1}^n (x_j - \theta_1)^2 + \sum_{j=1}^n (x_j - \theta_0)^2 \right\} \right] \\
 &= \exp \left[\frac{1}{2} \left\{ - \sum_{j=1}^n x_j^2 + 2\theta_1 \sum_{j=1}^n x_j - n\theta_1^2 + \sum_{j=1}^n x_j^2 - 2\theta_0 \sum_{j=1}^n x_j + n\theta_0^2 \right\} \right] \\
 &= \exp \left[(\theta_1 - \theta_0) \sum_{j=1}^n x_j \right] \exp \left[\frac{n}{2} (\theta_0^2 - \theta_1^2) \right]. \tag{2.203}
 \end{aligned}$$

In this case the critical region is

$$\exp \left[(\theta_1 - \theta_0) \sum_{i=1}^n x_i \right] \exp \left[\frac{n}{2} (\theta_0^2 - \theta_1^2) \right] \geq k. \tag{2.204}$$

Solving (2.204) using $(\theta_1 - \theta_0) > 0$ gives

$$\sum_{i=1}^n x_i \geq a, \quad a = \frac{\log \{ k \exp [-n (\theta_0^2 - \theta_1^2) / 2] \}}{(\theta_1 - \theta_0)}. \tag{2.205}$$

Thus, our test is to reject H_0 when $\sum_{i=1}^n x_i \geq a$, where a is chosen so that

$$P \left\{ \sum_{i=1}^n x_i \geq a \mid H_0 \right\} = \alpha. \tag{2.206}$$

Under H_0 , X_i is $N(\theta_0, 1)$ so that $(\sum_{i=1}^n x_i - n\theta_0) / \sqrt{n}$ is $N(0, 1)$. This gives

$$\begin{aligned}
 P \left\{ \sum_{i=1}^n x_i \geq a \mid H_0 \right\} &= P \left\{ \frac{\sum_{i=1}^n x_i - n\theta_0}{\sqrt{n}} \geq \frac{a - n\theta_0}{\sqrt{n}} \mid H_0 \right\} \\
 &= P \{ Z \geq (a - n\theta_0) / \sqrt{n} \}. \tag{2.207}
 \end{aligned}$$

By the Neyman-Pearson lemma, the best critical region is given by

$$(a - n\theta_0) / \sqrt{n} = z_\alpha \tag{2.208}$$

where $P \{ Z \geq z_\alpha \} = \alpha$. Thus we will reject H_0 if

$$\sum_{j=1}^n x_j \geq \sqrt{n} z_\alpha + n\theta_0, \tag{2.209}$$

or equivalently if $(\sum_{j=1}^n x_j - n\theta_0) / \sqrt{n} \geq z_\alpha$, where (x_1, x_2, \dots, x_n) are the observed outcomes.

As a numerical example, suppose we wish to test $\theta = 1$ against $\theta = 2$ for a sample of size 25 and $\alpha = 0.05$. Suppose $\sum_{j=1}^n x_j = 30$, what decision is made? By Eq. (2.209) we

will reject H_0 if $\sum_{j=1}^{25} x_j \geq \sqrt{25}z_{0.05} + 25 = 33.2$. Since $\sum_{j=1}^n x_j = 30 < 33.2$ we accept the hypothesis that $\theta = 1$ at the level of significance 0.05.

Before examining more general hypothesis testing problems we make some further comments about the Neyman-Pearson lemma. First, if the unknown parameter is a vector $\theta = (\theta_1, \theta_2, \dots, \theta_m)$, then Theorem 2.2 provides the best test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ since the scalar nature of θ plays no role in the proof. Second, the Neyman-Pearson lemma sometimes provides a best test for a simple hypothesis against a *composite alternative* in the sense that the critical region minimizes β simultaneously for all values $\theta_1 \in \Theta_1$.

For example in Example 2.29 we should note that the critical region does not depend on θ_1 as long as $\theta_1 > \theta_0$. Thus the critical region $\sum_{j=1}^n x_j \geq \sqrt{n}z_\alpha + n\theta_0$ minimizes β for all $\theta_1 > \theta_0$. Such tests are usually called *uniformly most powerful* (UMP) tests.

2.12.2 Generalized Likelihood Ratio Tests

Suppose now that we wish to devise good tests to test the general hypothesis $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. If either H_0 or H_1 is composite there is no generalization of the Neyman-Pearson lemma available. However, it seems reasonable that good tests may be devised by considering the likelihood ratio

$$L = f_{\mathbf{X}}(\mathbf{x}; \theta_1) / f_{\mathbf{X}}(\mathbf{x}; \theta_0), \quad \theta_0 \in \Theta_0, \theta_1 \in \Theta_1. \quad (2.210)$$

We will restrict our attention to the case of testing a simple hypothesis against the possibly, composite, alternative $H_1 : \theta \in \Theta_1$.

In this case $\theta \in \Theta_1$ is unknown and so must be estimated in some fashion. If we use the MLE $\hat{\theta}_1$ of θ then the analogue of the Neyman-Pearson lemma is to reject H_0 if

$$L = f_{\mathbf{X}}(\mathbf{x}; \hat{\theta}_1) / f_{\mathbf{X}}(\mathbf{x}; \theta_0) \geq k, \quad (2.211)$$

where k is chosen so that

$$P\{L \geq k | H_0\} = \alpha. \quad (2.212)$$

The test determined by (2.211)-(2.212) is called the *generalized likelihood ratio test* (GLRT) of H_0 against H_1 .

As a particular case consider the situation of testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, $\theta \in \mathbb{R}$. Here $\theta_1 = \mathbb{R} - \theta_0$, and in cases where X has a density it is permissible to find the MLE of θ over \mathbb{R} (\mathbb{R} = real numbers), since generally $P\{\hat{\theta}_1 = \theta_0\} = 0$.

Example 2.30 Let X be a $N(\theta, 1)$ random variable. Determine the generalized likelihood ratio test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

From Example 2.19 the maximum likelihood estimator of $\theta \in \mathbb{R}$ is $\hat{\theta} = \sum_{i=1}^n x_i / n$.

Thus

$$\begin{aligned}
 f_{\mathbf{X}}(\mathbf{x}; \hat{\theta}) &= \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right] \\
 &= \exp \left[-\frac{1}{2} \sum_{i=1}^n \{(x_i - \theta_0) - (\bar{x}_n - \theta_0)\}^2 \right] \\
 &= \exp \left[-\frac{1}{2} \left\{ \sum_{i=1}^n (x_i - \theta_0)^2 + n(\bar{x}_n - \theta_0)^2 \right\} \right]
 \end{aligned} \tag{2.213}$$

Using this in (2.211) gives

$$L = \exp \left[n(\bar{x}_n - \theta_0)^2 / 2 \right]. \tag{2.214}$$

Thus, the generalized likelihood ratio test is equivalent to rejecting H_0 if

$$(\bar{x}_n - \theta_0)^2 \geq a, \tag{2.215}$$

and this is equivalent to rejecting H_0 if either $\bar{x}_n - \theta_0 \geq \sqrt{a} = b$ or $\bar{x}_n - \theta_0 \leq -\sqrt{a} = -b$. (This test is usually called a *two-tailed test*.) b may be determined as in Example 2.29.

When H_0 is true we recognize the left hand side of (2.207) as the value of T^2 , where T is the T statistic discussed in Sections 2.8 and 2.11. The likelihood ratio test in this case becomes: reject H_0 if either

$$T \geq b \text{ or } T \leq -b \tag{2.216}$$

where b is chosen so that

$$P\{|T| \leq b\} = 1 - \alpha. \tag{2.217}$$

Since T has a t density with $n - 1$ degrees of freedom $b = t_{n-1, \alpha/2}$ and the critical region is

$$T \leq -t_{n-1, \alpha/2} \text{ or } T \geq t_{n-1, \alpha/2}. \tag{2.218}$$

2.13 Hypothesis Testing and Confidence Intervals

In Section 2.12 we observed that the t test for normal means was based on the same statistic used in developing a confidence interval for the same parameter in Section 2.11. This fact (and similar ones) leads one to examine that relationship between the notions of confidence intervals and hypothesis testing.

Suppose then that $(\hat{\theta}_L, \hat{\theta}_U)$ is an exact $(1 - \alpha) \times 100\%$ confidence interval for a parameter θ , then $\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$ with probability $1 - \alpha$. Since $\hat{\theta}_L$ and $\hat{\theta}_U$ place bounds on the possible values of θ we can use this information to develop a test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. If $\theta_0 \in (\hat{\theta}_L, \hat{\theta}_U)$ we accept the hypothesis, otherwise we reject it.

This gives a test whose critical region consists of the set $\mathcal{C} = \{\theta_0 > \hat{\theta}_U\} \cup \{\theta_0 < \hat{\theta}_L\}$. Since

$$P\{\mathcal{C}|H_0\} = 1 - P\{\hat{\theta}_L \leq \theta_0 \leq \hat{\theta}_U\} = 1 - (1 - \alpha) = \alpha, \tag{2.219}$$

we have a test of size α . (Sometimes such tests are equivalent to a generalized likelihood ratio test, such as the case of the t -test, other times they are not.)

One-sided confidence intervals may be used to develop tests of one-sided hypotheses. For example, suppose we have a *lower confidence limit* $\hat{\theta}_L$ where $P\{\hat{\theta}_L \leq \theta\} = 1 - \alpha$. Then, if we want to test the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ we can do it in the following way: accept H_0 if $\hat{\theta}_L \leq \theta_0$ and reject H_0 otherwise. In this case, we have a test of size α whose critical region is $\hat{\theta}_L \geq \theta_0$. As a particular case, suppose that X is $N(\theta, \sigma^2)$ where σ is known. Then $\hat{\theta}_L = \bar{X}_n - z_\alpha \sigma / \sqrt{n}$ is a $(1 - \alpha) \times 100\%$ lower confidence limit for θ . Then, if we apply the critical region determined by $\hat{\theta}_L$, the critical region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ is: reject H_0 if $\bar{X}_n - z_\alpha \sigma / \sqrt{n} > \theta_0$ or $\bar{X}_n \geq \theta_0 + z_\alpha \sigma / \sqrt{n}$. This is the likelihood ratio test obtained in Example 2.29 (there $\sigma = 1$).

A similar procedure may be used for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ by using an upper confidence limit $\hat{\theta}_U$ for θ . The details are left to the reader. This connection between confidence intervals and hypothesis testing will play a major role in our development of tests associated with various hypotheses in regression analysis.

2.14 Exercises

2.1 Consider an experiment that consists of tossing a fair coin and a 6-sided die.

- Describe the sample space with the elements for the experiment.
- What is the probability of the event that either a head of the coin occurs or the number of dots is greater than four?

2.2 For given two events A_1 and A_2 , find the union and the intersection where

- $A_1 = \{0, 1, 2, 3\}$, $A_2 = \{2, 3, 4, 5\}$.
- $A_1 = \{x : 0 \leq x < 2\}$, $A_2 = \{x : 1 < x < 3\}$.

2.3 Let A and B be two events such that $P(A) = 0.48$, $P(A \cup B) = 0.72$.

- Find $P(B)$ if $P(A \cap B) = 0.24$.
- Find $P(B)$ if $P(\overline{B}) = 0.64$.
- Find $P(B)$ if A, B are mutually exclusive.
- Find $P(B)$ if A, B are independent.
- Find $P(B)$ if $P(A|B) = 0.36$.

2.4 Let $f(x) = x/10$, $x = 1, 2, 3, 4$, zero elsewhere, be the pdf of a random variable X .

- Find the distribution function $F(x)$ of X and sketch its graph along with that of $f(x)$.
- Find $P(X = 1 \text{ or } 2)$ and $P(1 \leq X \leq 2)$.

2.5 Let the random variables X and Y have the joint pdf

$$f(x, y) = \begin{cases} 2, & 0 < x < y < 1, 0 < y < 1 \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Find the marginal probability distributions of X and Y , respectively.
 (b) Show that the correlation coefficient of X and Y , $\rho_{X,Y} = 1/2$.
 (c) Find the conditional distribution, $f(y|x)$ and the conditional mean, $E(Y|x)$.
 (d) If the conditional mean of Y , given $X = x$, is linear in y , then that conditional mean is given by

$$E(Y|x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X).$$

Verify this in (c).

- (e) Verify that the conditional variance of Y , given $X = x$, $Var(Y|x) = \sigma_Y^2 (1 - \rho_{X,Y}^2)$.

2.6 Suppose that the bivariate density for (X, Y) is given

$$f(x, y) = \begin{cases} 8xy, & 0 \leq x \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Find $Var(X + Y)$.
 (b) Find the coefficient of correlation, $\rho_{X,Y}$.

2.7 Show that $X + Y$ and $X - Y$ are uncorrelated if and only if $Var(X) = Var(Y)$.

2.8 Show that $Cov(X, X + Y) = Var(X) + Cov(X, Y)$. More generally, show that

$$Cov\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_i \sum_j a_i b_j Cov(X_i, Y_j).$$

2.9 Show that each of the following families of distributions is an exponential family.

- (a) normal distribution with either parameter μ or σ known; $X \sim N(\mu, \sigma^2)$.
 (b) gamma distribution with either parameter α or β known; $X \sim G(\alpha, \beta)$.

$$[f_X(x) = \beta^\alpha x^{\alpha-1} \exp(-\beta x) / \Gamma(\alpha).]$$

- (c) Poisson distribution with parameter λ ; $X \sim Poisson(\lambda)$.

2.10 Given that X have a Poisson distribution with variance 1, find $P(X = 2)$ and $P(X > 2)$.

2.11 Let the random variable X have a normal distribution with the mean μ and variance σ^2 . Consider a transformation $Z = (X - \mu) / \sigma$.

- (a) Show that $E(X) = 0$, and $E(Z^2) = 1$.
 (b) Let $M(t)$, $-h < t < h$ denote the moment generating function of the random variable X , show that $E[\exp(tZ)] = \exp(-\mu t / \sigma) M(t / \sigma)$, $-h\sigma < t < h\sigma$.

2.12 Let X be a random variable whose second moment exists.

- (a) Show that $E[(X - b)^2]$ is a minimum when $b = E(X)$ for all real number b .
 (b) Show that $E(X^2) \geq [E(X)]^2$.

2.13 Let X have a binomial distribution with the parameters n and p .

- (a) Find the moment generating function of X , $M_X(t)$.
- (b) Using the result in (a) find the mean and the variance of X .
- (c) Compute $E(X)$ and $Var(X)$ by direct summation.

2.14 If X is a Poisson random variable, calculate $P\{X = x | X \geq y\}$. (These distribution for fixed y and variable x is sometimes called *conditional Poisson distributions*.)

2.15 Suppose X follows a Poisson distribution with parameter λ . Then, is $P\{X \text{ takes an even value}\} = P\{X \text{ takes an odd value}\}$?

2.16 Let $\Phi(z)$ denote the cumulative distribution function of standard normal random variable Z , i.e., $\Phi(z) = P(Z \leq z)$. Show that $\Phi(-z) = 1 - \Phi(z)$.

2.17 If X is $N(\mu, \sigma^2)$, find c such that

- (a) $P\{-c < (X - \mu)/\sigma < c\} = 0.90$.
- (b) $P\{|(X - \mu)/\sigma| > c\} = 0.05$.

2.18 Suppose that X is normally distributed with mean $\mu = 10$ and variance $\sigma^2 = 4$.

- (a) Compute $P(|X - 9| > 1)$.
- (b) Find x such that $P(X > x) = 0.9750$.
- (c) Find the mean and the variance of $Y = 3X - 2$.

2.19 Suppose that X/σ^2 has a chi-square distribution with 10 degrees of freedom. Determine the pdf, mean, and variance of X .

2.20 Let X be $\chi^2(30)$.

- (a) Find $E(X)$ and $Var(X)$. Use these to approximate $P(18.5 < X < 43.8)$.
- (b) Find the exact value of $P(18.5 < X < 43.8)$.

2.21 Let s^2 be the variance of a random sample of size $n = 6$ from the normal distribution with unknown μ and variance $\sigma^2 = 12$. Find $P(2.30 < s^2 < 22.2)$.

2.22 Find the mean and variance of $s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, where X_1, X_2, \dots, X_n denote a random sample of size n from a $N(\mu, \sigma^2)$. [Hint: Find the mean and variance of $(n-1)s^2/\sigma^2$.]

2.23 Let U have a uniform distribution on $[0, 1]$. Show that the variable $-2\log U$ has a χ^2 -distribution with 2 degrees of freedom. [Hint: Consider $P(-2\log U \leq x)$.]

2.24 Suppose that T has a t -distribution with 8 degrees of freedom. Find $P(|T| > 2.306)$.

2.25 Let T have a t -distribution with 12 degrees of freedom. Find k such that

- (a) $P(-k < T < k) = 0.90$.
- (b) $P(|T| > k) = 0.99$.

2.26 Let F have an F -distribution with degrees of freedom ν_1 and ν_2 . Show that $1/F$ has an F -distribution with degrees of freedom ν_2 and ν_1 .

- 2.27** If F has an F -distribution with degrees of freedom $\nu_1 = 2$ and $\nu_2 = 6$. Find a and b such that $P(F \leq a) = 0.05$ and $P(F \leq b) = 0.95$, and accordingly, $P(a < F < b) = 0.90$.

[Hint: Write $P(F \leq a) = P(1/F \geq 1/a) = 1 - P(1/F \leq 1/a)$, and use the result of Exercise 2.26.]

- 2.28** Let a random variable X have a lognormal distribution and let Y have a normal distribution.

- Using the cdf technique, find the pdf of X . [See hint in Ex. 2.23.]
- Express the k -th central moment of X using mgf of $\log X$.
- Use the result in (b) find the mean and the variance of X .

- 2.29** Let θ denote a parameter and define the *bias* $B(\hat{\theta}) = |E(\hat{\theta}) - \theta|$. For an estimator $\hat{\theta}$, show that the mean square error of the point estimator $\hat{\theta}$,

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = Var(\hat{\theta}) + [B(\hat{\theta})]^2.$$

- 2.30** Let X_1, X_2, \dots, X_n be a random sample of size n from a Poisson distribution with mean λ . Find the method of moments estimator of λ .

- 2.31** Let X_1, X_2, \dots, X_n denote a random sample of size n from the normal distribution with mean $\mu = 0$ and unknown σ^2 . Find the method of moments estimator of σ^2 .

- 2.32** Suppose that X_1, X_2, \dots, X_n constitute a random sample from the gamma distribution with parameters (α, β) . Find the method of moments estimators of (α, β) . [Hint: Find $E(X)$, $Var(X)$.]

- 2.33** Let X_1, X_2, \dots, X_n be i.i.d. random variables with common pdf (or pmf). Find an MLE for θ in each of the following cases.

- $f(x) = (1/\theta) \exp(-x/\theta)$, $0 < x < \infty$, $\theta > 0$.
- $f(x) = \exp(-x + \theta)$, $\theta \leq x < \infty$.
- $p(x) = 1/\theta$, $x = 1, 2, \dots, \theta$, $1 \leq \theta \leq \theta_0$, θ_0 is a known integer.

- 2.34** Let X_1, X_2, \dots, X_n be a random sample of size n from a Poisson distribution with mean λ .

- Find the maximum likelihood estimator $\hat{\lambda}$ for λ .
- Find $E(\hat{\lambda})$ and $Var(\hat{\lambda})$.
- Show that the estimator in (a) is consistent for λ .
- Find the MLE for $P(X = 0) = \exp(-\lambda)$.

- 2.35** Suppose that we have a random sample of size 5, X_1, X_2, \dots, X_5 , from an exponential distribution with pdf given by $f(x) = (1/\theta) \exp(-x/\theta)$, $x > 0$. Suppose that we consider the following estimators;

$$\begin{aligned} \hat{\theta}_1 &= X_1, & \hat{\theta}_2 &= \frac{X_1 + X_4}{2} + 2, \\ \hat{\theta}_3 &= \frac{X_1 + X_3 + X_5}{3}, & \hat{\theta}_4 &= \bar{X}, \end{aligned}$$

where $\bar{X} = \frac{1}{5} \sum_{i=1}^5 X_i$.

- (a) Determine which estimators of θ are unbiased.
- (b) Calculate the variance of each of the estimators.
- (c) Calculate the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_3$, and of $\hat{\theta}_3$ relative to $\hat{\theta}_4$.

2.36 Let X_1, X_2, \dots, X_n be a random sample of size n from a certain uniform distribution. Find the MLE of θ if

- (a) the distribution is $U(0, \theta)$.
- (a) the distribution is $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

2.37 Let X_1, X_2, \dots, X_n denote a random sample of size n from the Poisson distribution with parameter λ . Show that $\sum_{i=1}^n X_i$ is sufficient for λ .

2.38 Use the factorization criterion to determine, in each case, a sufficient statistic based on a random sample size of n .

- (a) $p(x) = p^x (1-p)^{1-x}$, $x = 0, 1$.
- (b) $p(x) = p(1-p)^x$, $x = 0, 1, 2, \dots$
- (c) $f(x) = \lambda \exp(-\lambda x)$, $x > 0$.

2.39 Let the observed value of the mean \bar{Y} of a random sample of size 15 from a $N(\theta, 52)$ be 13.2. Construct a 90% confidence interval for θ .

2.40 Let X_1, X_2, \dots, X_{10} denote a random sample of size $n = 10$ from a $N(\mu, \sigma^2)$ random variable.

- (a) If σ is known, find the length of a 95% confidence interval for μ if this interval is based on the random variable $\sqrt{10}(\bar{X} - \mu)/\sigma$.
- (b) If σ is unknown, find the expected value of the length of a 95% confidence interval for μ if this interval is based on the random variable $\sqrt{9}(\bar{X} - \mu)/s$.

2.41 Suppose that X_1, X_2, \dots, X_{200} is a random sample from a Bernoulli(p) population. Let $\hat{p} = \sum x_i/n$ and if the observed value of $\hat{p} = 0.25$, find an approximate 90 confidence interval for the true proportion p .

2.42 Let X_1, X_2, \dots, X_{16} be a random sample of size $n = 16$ from a $N(\mu, 1)$ random variable. We wish to test $H_0 : \mu = 20$ against $H_0 : \mu \neq 20$ at $\alpha = 0.05$, based on the sample mean \bar{X} .

- (a) Determine critical regions of the form $C_1 = \{\bar{x} : \bar{x} \leq a\}$ and $C_2 = \{\bar{x} : \bar{x} \geq b\}$.
- (b) Find β , the probability of a Type II error, for each critical region in (a) for the alternative $H_1 : \mu = 21$.

2.43 Let X_1, X_2, \dots, X_{25} denote a random sample of size $n = 25$ from a $N(\mu, 100)$ random variable. Derive the uniformly most powerful (UMP) test with size $\alpha = 0.10$ for testing $H_0 : \mu = 75$ against $H_1 : \mu > 75$.

2.44 Let X_1, X_2, \dots, X_n denote a random sample from the Poisson distribution with parameter θ .

(a) Find a UMP test with size $\alpha = 0.05$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

(b) Sketch the power function, $\kappa(\theta)$ for $\theta_0 = 1$ and $n = 25$. Take $\alpha = 0.05$. [Hint: Use the Central Limit Theorem.]

2.45 Find a generalized likelihood ratio test of size α for testing $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$ based on a random sample X_1, X_2, \dots, X_n from $f(x) = \theta \exp(-\theta x)$, $0 < x < \infty$.

Chapter 3

Simple Linear Regression

3.1 Introduction

In this chapter we begin our formal analysis of regression models by considering the *simple linear regression model*

$$Y_x = \beta_0 + \beta_1 x + \varepsilon_x. \quad (3.1)$$

Here, Y_x is the *dependent* or *response variable*, x is the *independent* or *design variable* and ε_x is an error random variable used to represent the variation of Y_x not explained by the linear part $\beta_0 + \beta_1 x$. Conceptually, we may think of the values of Y_x as generated by choosing a value of x , computing $\beta_0 + \beta_1 x$ and adding the random error given by ε_x . Thus, if we pick n values of $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and plot the resulting data $(x_i, y_i \equiv y_{x_i})$, then depending on the variability of ε_x , these points should scatter around the line $y = \beta_0 + \beta_1 x$. Of course, in practice, only the data (x_i, y_i) , $1 \leq i \leq n$, will be known and the purpose of the statistical analysis is to determine if the model (3.1) is capable of explaining the observed variability in the y 's. In addition, since some of $(\beta_0, \beta_1, \varepsilon_x)$ are usually unknown, our first task is to find estimates of these quantities. Comparison of the fitted model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad (3.2)$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ are estimates of (β_0, β_1) , with the known data then provides a way of determining whether or not (3.1) appears to be appropriate. Although many techniques are available for doing this, for instance, merely drawing an 'eyeball' line through $\{(x_i, y_i)\}_{i=1}^n$, consideration of more elementary statistical situations, knowing the distribution of Y_x , and hence of ε_x , is crucial. However, a priori, there is no more reason for knowing this than knowing (β_0, β_1) and this argument seems to bring us to a logical impasse. To proceed further there are two avenues of approach:

- (1) estimate (β_0, β_1) using a method which does not depend on the distribution of Y_x ;
- (2) make a plausible assumption about the error structure, estimate (β_0, β_1) and then check the assumptions.

As we shall see, in the important case where the errors ε_x are assumed to have a normal distribution and *maximum likelihood estimation* is used to obtain $(\hat{\beta}_0, \hat{\beta}_1)$ we obtain a method of estimation, *least squares*, which can be applied to studying linear models regardless of the underlying error distribution. This technique, apparently first derived by Gauss in the 18th century, is the most widely used method of estimation in regression analysis and most of this book will be devoted to analyzing its consequences. However, because of problems which can occur when the errors are not normal, much work has been done recently on alternative estimation procedures, such as maximum likelihood and robust regression technique. Some of these are discussed in [27, 87].

3.2 The Error Model

As we have previously stated, the simple linear regression model is of the form $Y_x = \beta_0 + \beta_1 x + \varepsilon_x$ where we now assume that ε_x does not depend on x and $E(\varepsilon_x) = 0$ so that

$$E(Y_x) = \beta_0 + \beta_1 x. \quad (3.3)$$

If n observations of Y_x are made at the points $\{x_i\}_{i=1}^n \equiv \{x\}_1^n$, then letting $Y_i \equiv Y_{x_i}$ denote the i -th observation,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (3.4)$$

where $\varepsilon_i \equiv \varepsilon_{x_i}$, $i = 1, 2, \dots, n$.

To estimate (β_0, β_1) we assume that Y_i , $1 \leq i \leq n$ are *independent normal random variables* with mean $\beta_0 + \beta_1 x_i$ and common variance σ^2 . This is equivalent to assuming that $\{\varepsilon_i\}_1^n$ are normal random variables with mean zero and variance σ^2 (This is usually written as $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and $\varepsilon_i \sim N(0, \sigma^2)$.) In this case we say that the errors are *homoscedastic* (otherwise they are *heteroscedastic*).

Since maximum likelihood estimation is known to be optimal for estimating the mean μ of a normal random variable, in that it gives $\hat{\mu} = \bar{x}$, the sample mean, which is the *minimum variance unbiased estimator* of μ , it is reasonable to consider this approach for the estimation of (β_0, β_1) in (3.4).

Now if

$$\frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 \right] \quad (3.5)$$

is the density of Y_i , then because $\{Y_i\}_1^n$ are assumed to be independent random variables, the likelihood function L is given by

$$\begin{aligned} L &= \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 \right] \right\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right] \end{aligned} \quad (3.6)$$

As usual, the maximum likelihood estimates of (β_0, β_1) are obtained by maximizing L with respect to (β_0, β_1) . Since L is a “negative exponential”, to maximize it it suffices to minimize the exponent

$$\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (3.7)$$

in (3.7). Since $\sigma^2 > 0$, it suffices to minimize

$$S = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (3.8)$$

This can be done conveniently using calculus methods and will be pursued here. An alternative algebraic proof is given in Section 3.4.

Taking the partial derivatives of S with respect to (β_0, β_1) and setting these derivatives to zero enables us to find the minimizing values. From (3.8)

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i), \quad (3.9)$$

and

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i. \quad (3.10)$$

Setting the right hand side of (3.9) to zero and rearranging gives

$$n\hat{\beta}_0 = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \quad (3.11)$$

where the “carats” over (β_0, β_1) indicate that these are the estimated values.

Solving (3.11) for $\hat{\beta}_0$ gives

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3.12)$$

where we have used the notation

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (3.13)$$

To obtain an expression for $\hat{\beta}_1$ we set (3.10) to zero giving

$$\sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0, \quad (3.14)$$

and substituting (3.12) for $\hat{\beta}_0$ in (3.14) gives

$$\sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0. \quad (3.15)$$

Thus,

$$\left(n\bar{x}^2 - \sum_{i=1}^n x_i^2 \right) \hat{\beta}_1 = n\bar{x}\bar{y} - \sum_{i=1}^n x_i y_i. \quad (3.16)$$

Using the additional notation

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2 \quad \text{and} \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i, \quad (3.17)$$

and solving (3.16) for $\hat{\beta}_1$ gives

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (3.18)$$

Having obtained $\hat{\beta}_1$ from (3.18) we may then calculate $\hat{\beta}_0$ from

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \quad (3.19)$$

with $\hat{\beta}_1$ given by (3.18).

Once $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined from (3.18)-(3.19), we can obtain estimates of the y values from the fitted line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \quad (3.20)$$

In particular, we can estimate $E(Y_i)$ by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \quad (3.21)$$

The difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad (3.22)$$

is called the i -th *residual* and its size is an indicator of how well the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ fits the observed data. The residuals may also be viewed as estimates of ε_i , and their examination plays an important role in assessing the reasonableness of the assumptions given for (3.1).

Before proceeding with the estimation of σ^2 we make several comments concerning the MLEs of (β_0, β_1) . First, even if the errors are not normal, we can still estimate (β_0, β_1) by minimizing S which is a sensible procedure if we consider the geometric interpretation of this process. Suppose that $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is the equation of some line which we estimate to fit the “true line” $y = \beta_0 + \beta_1 x$. Then $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ represents the vertical deviation of the point (x_i, y_i) from this estimated line and

$$S = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \quad (3.23)$$

is a measure of how well the estimated line fits the data. It is reasonable to choose as a candidate for the best line one that makes (3.23) the smallest. If we choose $(\hat{\beta}_0, \hat{\beta}_1)$ by minimizing (3.23) to do this, we will arrive at (3.18)-(3.19) as our estimates of the *intercept* β_0 and *slope* β_1 , in agreement with the MLEs when the errors are normal. In this case $(\hat{\beta}_0, \hat{\beta}_1)$ are called the *least squares estimators* and the resultant fitted line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ the *least squares line*. Since many measures of goodness of fit other than (3.23) are possible, one justification for using the least squares estimators for arbitrary error distributions comes from the fact that when the errors are normal the least squares and maximum likelihood estimators agree. For other error distributions they will not.

As an example, suppose that the density of ε_i is given by

$$f_{\varepsilon_i}(\varepsilon) = \frac{1}{2} \exp(-|\varepsilon|). \quad (3.24)$$

In this case ε_i has a *Laplace* or *double exponential distribution*. If the errors are independent and identically distributed according to (3.24), then the density of Y_i is given by

$$f_{Y_i}(y_i) = \frac{1}{2} \exp(-|y_i - \beta_0 - \beta_1 x_i|). \quad (3.25)$$

and the likelihood function for n independent observations is given by

$$L = \frac{1}{2^n} \exp\left(-\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|\right). \quad (3.26)$$

In this case the MLE of (β_0, β_1) is given by minimizing

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \quad (3.27)$$

and the values of $(\hat{\beta}_0, \hat{\beta}_1)$ minimizing (3.27) will in most circumstances differ from those given by (3.18)-(3.19). As a numerical example, consider the three points $(0, 0)$, $(1, 0)$ and $(1/2, 1/2)$ shown in Figure 3.1. The least squares line is easily found to be $y = 1/6$. For this line the value of (3.23) is $2/3$ and so does not minimize (3.27), since for the line $y = 0$ the value is $1/2$. In fact, the values $\hat{\beta}_0 = \hat{\beta}_1 = 0$ minimize (3.27) in this case.

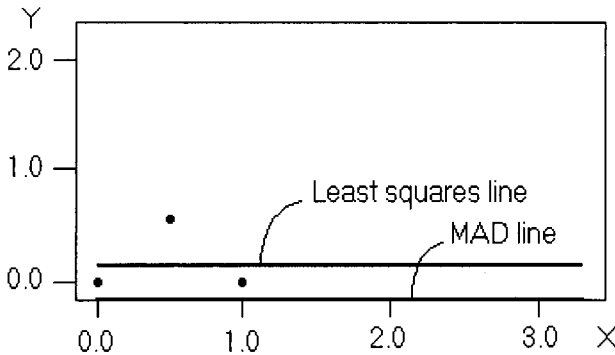


Figure 3.1: Least squares and Minimum Absolute Deviations (MAD) lines

On the other hand the least squares estimator is the *best linear unbiased estimator* (BLUE) (see Section 3.6), and this fact, the *Gauss-Markov theorem* is often invoked to justify its use even when the errors are not normal. For the most part in this text we will concentrate on least squares estimation.

Although we have assumed that the x 's are known exactly, this may not always be the case. Measuring instruments are not 100% reliable and in practice values of both (x, y) are often rounded to some convenient value. In such "*errors in variable*" (EIV) models other estimators are often preferred to the least squares ones. A popular technique for doing this is to minimize the sum of the perpendicular distances of the data points from

the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ with the coefficients $(\hat{\beta}_0, \hat{\beta}_1)$, now being determined by minimizing

$$\sum_{i=1}^n d_i^2 \quad (3.28)$$

where $d_i, 1 \leq i \leq n$ are shown in Figure 3.2. This technique is called *orthogonal least squares* in the statistics literature. This distance measure does not favor the x variable, as does ordinary least squares, but rather treats both x and y variables equally. It is also known as the method of *total least squares* in numerical analysis. It also needs to be noted that ordinary least squares has some problems in EIV models [69, 13].

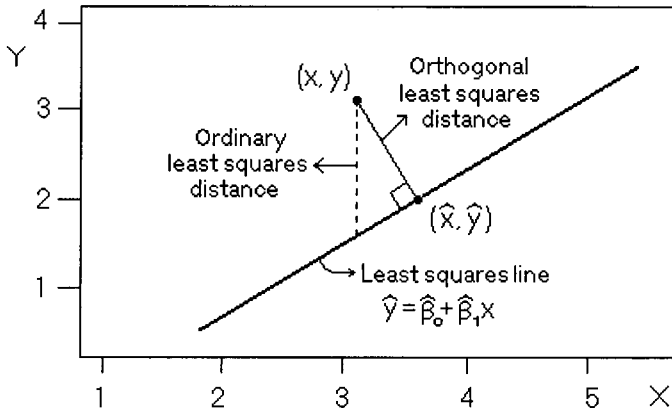


Figure 3.2: Distance minimized by orthogonal least squares

As shown in Figure 3.2, for a particular data point (x, y) , the point (\hat{x}, \hat{y}) on a line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ that is closest when we measure distance perpendicularly is given by (See Exercise 3.4)

$$\hat{x} = \frac{\hat{\beta}_1 y + x - \hat{\beta}_0 \hat{\beta}_1}{1 + \hat{\beta}_1^2}, \quad \hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{1 + \hat{\beta}_1^2} (\hat{\beta}_1 y + x - \hat{\beta}_0 \hat{\beta}_1). \quad (3.29)$$

3.2.1 Algebraic Derivation of the Least Squares Estimators

For those readers who have not had calculus, or for those who wish to see a purely algebraic derivation of the least squares estimators we supply one here.

First consider

$$\begin{aligned} y_i - \beta_0 - \beta_1 x_i &= y_i - \bar{y} - \beta_0 - \beta_1 (x_i - \bar{x}) + \bar{y} - \beta_1 \bar{x} \\ &= (y_i - \bar{y}) - \beta_1 (x_i - \bar{x}) + (\bar{y} - \beta_1 \bar{x} - \beta_0) \end{aligned} \quad (3.30)$$

where $i = 1, 2, \dots, n$.

Squaring each of the n terms given by (3.30) and adding yields

$$\begin{aligned} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\beta_0 - \bar{y} + \beta_1 \bar{x})^2 \\ &\quad - 2\beta_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - 2 \sum_{i=1}^n (y_i - \bar{y})(\beta_0 - \bar{y} + \beta_1 \bar{x}) \\ &\quad + 2\beta_1 \sum_{i=1}^n (x_i - \bar{x})(\beta_0 - \bar{y} + \beta_1 \bar{x}) \end{aligned} \quad (3.31a)$$

$$\begin{aligned} &= \sum_{i=1}^n (y_i - \bar{y})^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 + n(\beta_0 - \bar{y} + \beta_1 \bar{x})^2 \\ &\quad - 2\beta_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \end{aligned} \quad (3.31b)$$

where the last two terms in (3.31a) are zero because $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (y_i - \bar{y}) = 0$.

Now consider the term in (3.31b)

$$\beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2\beta_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \quad (3.32)$$

Completing the square in (3.32) it becomes

$$\begin{aligned} &\sum_{i=1}^n (x_i - \bar{x})^2 \left[\beta_1^2 - 2\beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right. \\ &\quad \left. + \left\{ \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^2 \right] - \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2, \end{aligned} \quad (3.33)$$

and using this in (3.31b) we get

$$\begin{aligned} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 - \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \\ &\quad + n(\beta_0 - \bar{y} + \beta_1 \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x})^2 \\ &\quad \cdot \left[\beta_1 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \end{aligned} \quad (3.34)$$

Since the first two terms in (3.34) do not depend on (β_0, β_1) , the sum of squares of the residuals will be minimized by setting the last two terms to zero gives

$$(\beta_0 - \bar{y} + \beta_1 \bar{x})^2 = 0, \quad (3.35)$$

and

$$\left[\beta_1 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 = 0. \quad (3.36)$$

Solving (3.35) and (3.36) gives

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (3.37)$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (3.38)$$

which agree with (3.18) and (3.19).

Note that we have obtained a little more with this argument than using calculus, since (3.34) shows that $(\hat{\beta}_0, \hat{\beta}_1)$ actually minimize (3.8), rather than just providing a critical point. We will expand on this approach in Chapter 5 when considering multiple regression.

As we have already pointed out, most regression calculations of any significance usually require a computer to do. However, in order to get a “feel” for the nature of the calculations and the basic concepts one should probably do at least some hand computations. (Pain and suffering are good for the soul.) For this purpose, we have chosen as our first example a simple three point model of no particular importance. More problems with computer generated results will follow.

Example 3.1 Find the MLEs for (β_0, β_1) in (3.1) when (x_i, y_i) have the values $(1, 1)$, $(2, 4)$, $(5, 7)$ and the errors are assumed to be independent $N(0, \sigma^2)$.

From our previous discussion the MLEs of (β_0, β_1) are given by (3.18)-(3.19). To obtain these values we use a classical computing format illustrated in Table 3.1.

Table 3.1 Calculations			
x	y	x^2	xy
1	1	1	1
2	4	4	8
5	7	25	35
<hr/>			
$\sum x = 8, \bar{x} = \frac{8}{3}$	$\bar{y} = 4$	$\overline{x^2} = 10$	$\overline{xy} = \frac{44}{3}$

In each column we list the values indicated. x -values of x_i ; y -values of y_i ; x^2 -values of x_i^2 ; and xy -values of $x_i y_i$. The last number in each column is the average of the corresponding column values. The various averages are then used in (3.18)-(3.19) to give $(\hat{\beta}_0, \hat{\beta}_1)$.

In this case we have

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{44/3 - 32/3}{10 - 64/9} = \frac{36}{26} = 1.3846154 \quad (3.39)$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4 - \frac{96}{26} = 0.3076923. \quad (3.40)$$

Although (3.18) is easy to remember due to the symmetric appearance of \overline{xy} , $\bar{x}\bar{y}$ etc. it is not always the best formula to use for numerical purposes.

For one thing, it is quite prone to round-off error and for machine calculations the equivalent form (left as an exercise)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.41)$$

is perhaps the better choice.

For theoretical purposes the expression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (3.42)$$

obtained from (3.41) by using the fact that $\sum_{i=1}^n \bar{y} (x_i - \bar{x}) = 0$, is often preferred. The following example illustrates the use of (3.42) numerically.

Example 3.2 An experiment was conducted to determine the relationship between the percentage of a certain drug in the blood stream and the reaction time to a given stimulus. The results are shown in Table 3.2.

Table 3.2 Drug and Reaction Time data		
Subject No.	Amount of drug x , %	Reaction time y , seconds
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

The scatter plot shown in Figure 3.3 suggests a possible linear relation between y and x . Assuming that the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, 5$$

where $\{\varepsilon_i\}_{i=1}^5$ are independent $N(0, \sigma^2)$ random variables is appropriate, find the MLEs of (β_0, β_1) .

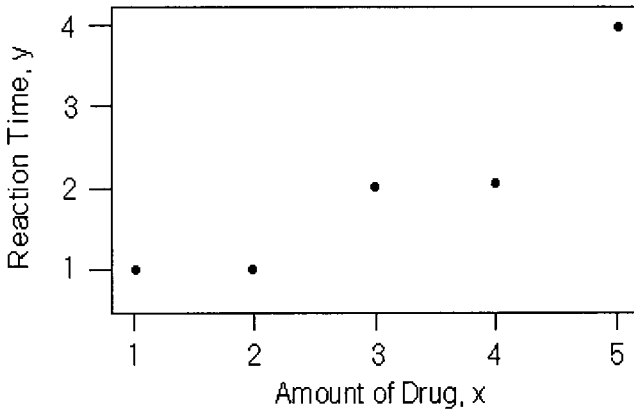


Figure 3.3: Scatter plot for drug and reaction time data

Since the ε_i 's are normal, the MLEs of (β_0, β_1) are the least squares estimates. To

obtain these we use the following tabular set-up similar to that in Example 3.1.

Table 3.3 Calculations for Drug and Reaction Time Data				
x	y	$x - \bar{x}$	$(x - \bar{x})^2$	$y(x - \bar{x})$
1	1	-2	4	-2
2	1	-1	1	-1
3	2	0	0	0
4	2	1	1	2
5	4	2	4	8
$\bar{x} = 3$	$\bar{y} = 2$	$\sum (x - \bar{x}) = 0$	$\sum (x - \bar{x})^2 = 10$	$\sum y(x - \bar{x}) = 7$

From this we find that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^5 y_i (x_i - \bar{x})}{\sum_{i=1}^5 (x_i - \bar{x})^2} = \frac{7}{10} = 0.7$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - 0.7(3) = -0.1.$$

The fitted line is then

$$\hat{y} = -0.1 + 0.7x.$$

A graph of this line superimposed on the scatter plot is shown in Figure 3.4.

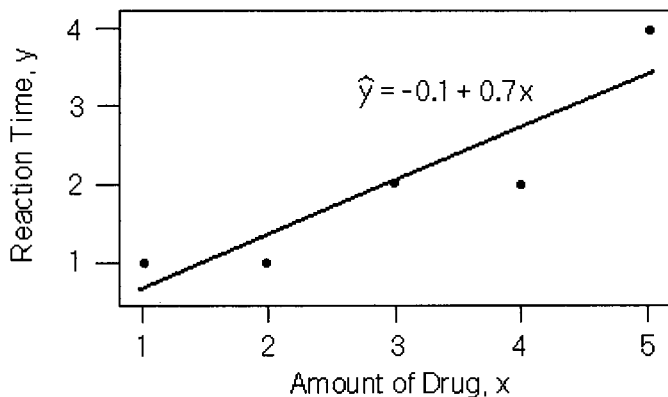


Figure 3.4: Fitted line for drug and reaction time data

We will now look at several more complicated examples of simple linear regression models. At this point we caution the reader to take our model assumptions concerning empirical data with a grain of salt. In general, one can never prove conclusively that the assumptions used to fit some data are ‘true’. The best one can hope to do is to make a fit given our assumptions, and then use a variety of statistical techniques, to see if our results appear to be consistent with these assumptions. If we feel they are, then we can entertain the hypothesized model as reasonable, otherwise we must modify our assumptions, refit and continue until we are satisfied. This process is only partly mathematical and different analysts may come to different conclusions for the same data. But we must start somewhere and (3.1) is often a good choice.

Example 3.3 To see what (3.18)-(3.19) will do when (3.1) is known to be true, we consider the following example taken from [40].

There five independent observations were made of a $N(0, 1)$ random variable and the values added to the five points (1, 8), (2, 11), (3, 14), (4, 17), (5, 20) on the line $y = 5 + 3x$. These observations gave rise to the data in Table 3.4.

Table 3.4 Observations (x, y)	
x	y
1	7.695
2	10.679
3	15.900
4	16.222
5	20.617

In this case the reader can verify that $\hat{\beta}_0 = 4.81$ and $\hat{\beta}_1 = 3.14$, showing that the least squares line provides a reasonable estimate of the true line, even for a small sample size.

For visual comparison the true line, estimated line and observed y values are shown in Figure 3.5.

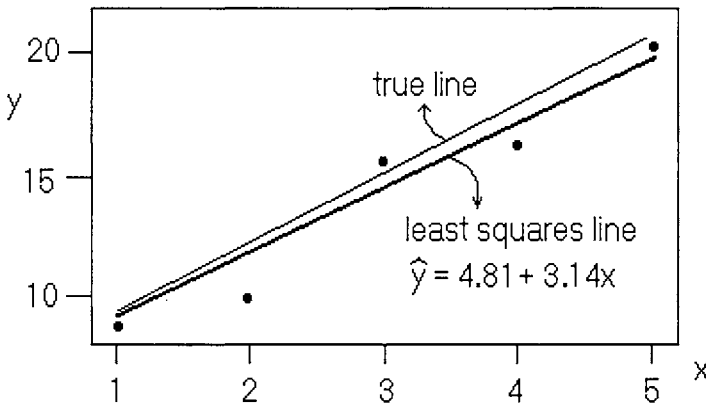


Figure 3.5: Comparison of true and fitted line for Example 3.3

Example 3.4 (Drink delivery data) To illustrate the further use of regression analysis in developing an empirical model, consider the following problem taken from [87].

A drink manufacturer is interested in determining if a linear relationship exists between the time y (in minutes) it takes to deliver an order and the number of cases delivered, x , is reasonable. For this he has available the data in Table 3.5. A scatter plot of these data is shown in Figure 3.6 and perhaps with the exception of the 9th observation the points appear to fall along a straight line. Thus if (3.1) is the true model, (we don't know if it is) then β_0 and β_1 may be estimated by least squares.

Table 3.5 The Number of Cases x and Delivery Time y

No.	Cases	Delivery time	No.	Cases	Delivery time
1	7	16.68	14	6	19.75
2	3	11.50	15	9	24.00
3	3	12.03	16	10	29.00
4	4	14.88	17	6	15.35
5	6	13.75	18	7	19.00
6	7	18.11	19	3	9.50
7	2	8.00	20	17	35.10
8	7	17.83	21	10	17.90
9	30	79.24	22	26	52.32
10	5	21.50	23	9	18.75
11	16	40.33	24	8	19.83
12	10	21.00	25	4	10.75
13	4	13.50			

Since the calculations here are tedious, values of $\hat{\beta}_0$ and $\hat{\beta}_1$ were obtained (by using MINITAB or S-PLUS). They are

$$\hat{\beta}_0 = 3.207798, \hat{\beta}_1 = 2.1761668$$

and the estimated line is shown in Figure 3.7 superimposed over the scatter plot. Without any formal assessment, the fit appears to be good and so at this point (3.1) seems to be a reasonable model of the given data.

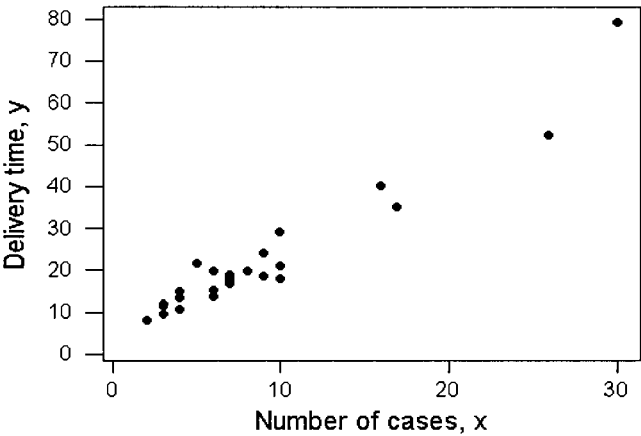


Figure 3.6: Scatter plot for delivery data

Example 3.5 (Tractor Data) The cost of maintaining a tractor in a given year appears to increase linearly with the age of the tractor. The following data were collected to examine this hypothesis. If a linear model is appropriate, find the least squares

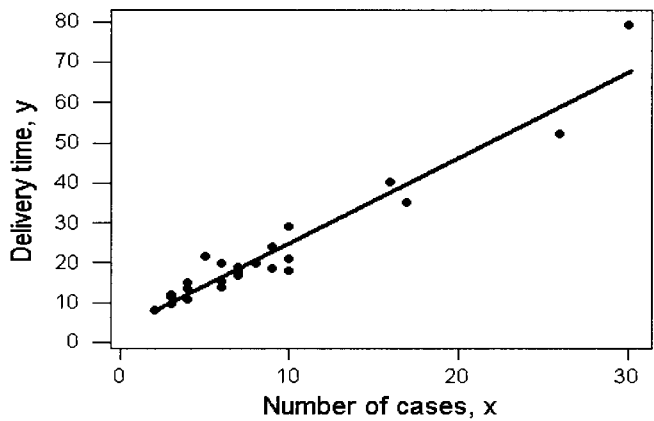


Figure 3.7: Least squares line for delivery data

estimates of (β_0, β_1) .

Table 3.6 Ages of Tractors and Maintenance Cost					
Obs. No.	Age x (years)	Cost y (\$)	Obs. No.	Age x (years)	Cost y (\$)
1	4.5	619	10	5.0	1,194
2	4.5	1,049	11	0.5	163
3	4.5	1,033	12	0.5	182
4	4.0	495	13	6.0	764
5	4.0	723	14	6.0	1,373
6	4.0	681	15	1.0	978
7	5.0	890	16	1.0	466
8	5.0	1,522	17	1.0	549
9	5.5	987			

Using (3.18)-(3.19) we find that

$$\hat{\beta}_0 = 323.622 \text{ and } \hat{\beta}_1 = 131.716$$

and so the fitted least squares line is

$$\hat{y} = 323.622 + 131.716x.$$

This line suggests that maintenance costs appear to increase at a rate of about \$131.72 per year.

Example 3.6 (Birth weight data) It is of interest to determine how the weight of a newborn baby depends on the length of the gestation period. To obtain a relationship the weights of 24 babies were measured (in grams) and the length of the corresponding gestation period (in weeks) was recorded. A scatter plot of the data is shown in Figure

3.8. A linear trend is observed so a line was fitted by least squares. (In fact, the plot looks more like two parallel lines.) The values of $(\hat{\beta}_0, \hat{\beta}_1)$ are given by

$$\hat{\beta}_0 = -1393 \text{ and } \hat{\beta}_1 = 113.20.$$

The result of this calculation indicates that the weight of a baby increases by about 113 grams for each week in the womb.

Table 3.7 Gestation Period and Weights of Newborn Babies

Obs. No.	Age (x)	Weight (y)	Obs. No.	Age (x)	Weight (y)
1	40	2968	13	40	3317
2	38	2795	14	36	2729
3	40	3163	15	40	2935
4	35	2975	16	38	2754
5	36	2625	17	42	3210
6	37	2847	18	39	2817
7	41	3292	19	40	3126
8	40	3473	20	37	2539
9	37	2628	21	36	2412
10	38	3176	22	38	2991
11	40	3421	23	39	2875
12	38	2975	24	40	3231

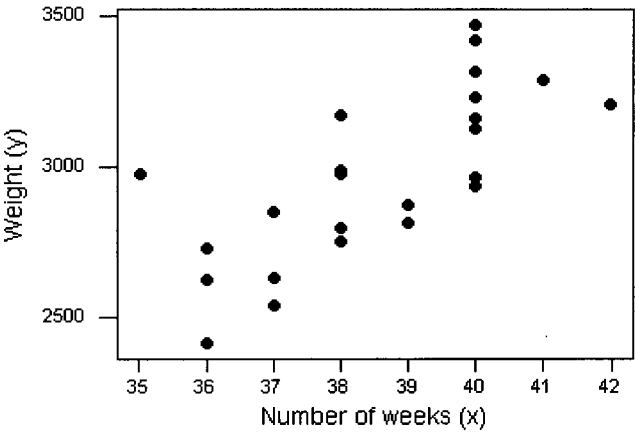
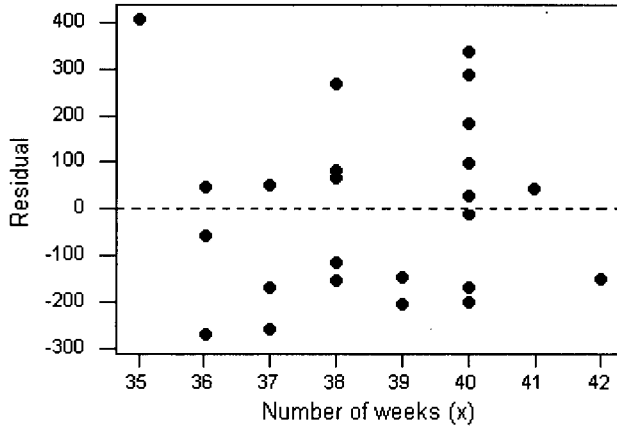


Figure 3.8: Scatter plot of weight and weeks for birth weight data

After we found the estimated linear regression model between the gestation period and weights of the twenty-four newborn babies, $y = -1393 + 113.20x$, we calculated the residuals, $\hat{\epsilon}_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, 24$, and plotted them in Figure 3.9.

As we shall see, the residual plot provides us with important clues for checking the assumptions about the postulated model. We will discuss this in more detail in Section 3.9.

Figure 3.9: Plot of residuals $\hat{\varepsilon} = y - \hat{y}$ for birth weight data

3.3 Estimating σ^2

As for (β_0, β_1) the model error σ^2 in (3.1) can be estimated by maximum likelihood. This can be obtained by differentiating L with respect to σ^2 . The result of this (see Exercise 3.5) calculation shows that

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (3.43)$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Thus the MLE of σ^2 is given by taking the average of the sum of squares of the residuals; an intuitive choice. By taking the square root of σ_{MLE}^2 we obtain the MLE of σ ,

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (3.44)$$

In this case the MLE is not the customary estimate of σ^2 ; rather the usual estimate of σ^2 is given by

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.45)$$

and σ is then estimated by

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (3.46)$$

which is usually called the *standard error of regression*.

The reason for choosing s^2 rather than $\hat{\sigma}_{MLE}^2$ to estimate σ^2 is that s^2 provides an *unbiased estimate* of σ^2 , regardless of the distribution of ε_i , $1 \leq i \leq n$, whereas, σ_{MLE}^2

is biased, even for normal errors. (However, s is a biased estimate of σ .) On average, $\hat{\sigma}_{MLE}^2$ underestimates σ^2 .

Using $n - 2$ in the denominator rather than n , is analogous to dividing by $n - 1$ rather than n in the usual estimate of σ^2 for random samples. In each instance one loses a “number of degrees of freedom” equal to the number of estimated parameters. The unbiasedness of s^2 will be demonstrated in the following section.

Although most regression calculations are done using calculators and/or computers it is possible to calculate s^2 without forming all the residuals $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, $1 \leq i \leq n$. This can be done according to the formula

$$s^2 = \frac{1}{n-2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right]. \quad (3.47)$$

This usually results in somewhat less computation than using (3.45) directly. To obtain (3.47) consider

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2 &= \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - x_i \hat{\beta}_1)^2 \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned} \quad (3.48)$$

Using

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.49)$$

(3.48) becomes

$$\sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2. \quad (3.50)$$

As a further matter of notation the numerator in the expression for s^2 will often be written as SSE ; short for ‘sum of the squares of the errors’. Then

$$s^2 = SSE / (n - 2). \quad (3.51)$$

Example 3.7 Assuming the errors in Example 3.3 have constant variance σ^2 , find an estimate of σ^2 by using s^2 . Here to illustrate (3.45) and (3.47) we will calculate s^2 in two ways.

First, using (3.45), $s^2 = SSE / (n - 2)$, so we need to obtain SSE . The relevant calculations are given below.

y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
1	0.6	0.4	0.16
1	1.3	-0.3	0.09
2	2.0	0	0
2	2.7	-0.7	0.49
4	3.4	0.6	0.36

$$SSE = 0.16 + 0.09 + 0 + 0.49 + 0.36 = 1.1.$$

Thus,

$$s^2 = 1.1/3 = 0.3667 \quad \text{and} \quad s = 0.6056.$$

To use (3.45), first we need to calculate

$$\sum_{i=1}^5 (y_i - \bar{y})^2 = (-1)^2 + (-1)^2 + (0)^2 + (0)^2 + (2)^2 = 6.$$

Then,

$$s^2 = \frac{1}{3} \left[\sum_{i=1}^5 (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^5 (x_i - \bar{x})^2 \right] = \frac{1}{3} [6 - (0.7)^2 10] = 0.3667$$

and $s = 0.6055$ as before.

Example 3.8 Calculate the standard error of regression for each of Examples 3.3, 3.4 and 3.5.

Here, because the calculations are tedious we do the calculations using MINITAB or S-PLUS. The results are:

- (1) for Example 3.3, $s = 1.204$;
- (2) for Example 3.4, $s = 4.181$;
- (3) for Example 3.5, $s = 283.479$.

We noted above that the reason usually given for using s^2 rather than the MLE of σ^2 is that s^2 is unbiased, while $\hat{\sigma}_{MLE}^2$ is not. However, s is generally a biased estimate of σ (as is $\hat{\sigma}_{MLE}$). As we now show this is a general property of unbiased estimators of the variance.

Let s^2 be an unbiased estimate of σ^2 . Then $E(s) \leq \sigma$. To see this, note that for any random variable X with $\sigma^2(X) < \infty$, that

$$E(X^2) = \text{Var}(X) + [E(X)]^2. \quad (3.52)$$

Taking $X = s$ in this equation gives

$$E(s^2) = \text{Var}(s) + [E(s)]^2 \quad (3.53)$$

But $E(s^2) = \sigma^2$, so that

$$\sigma^2 \geq [E(s)]^2 \quad (3.54)$$

or $E(s) \leq \sigma$, with equality holding only if $\text{Var}(s) = 0$. This occurs only if $P\{s^2 = 0\} = 1$. For normal random variables (see Chapter 2) s^2 is proportional to a χ^2 random variable with one degree of freedom. Hence, for normal random variables, $E(s) < \sigma$.

3.4 Properties of $(\hat{\beta}_0, \hat{\beta}_1, s^2)$

We now consider some properties of $(\hat{\beta}_0, \hat{\beta}_1)$ and of the estimator of variance s^2 . These properties will be important in developing confidence intervals and significance tests for the parameters.

Theorem 3.1 *Let $(\hat{\beta}_0, \hat{\beta}_1)$ be the least squares estimators of (β_0, β_1) in the linear model*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad 1 \leq i \leq n, \quad (3.55)$$

where $\varepsilon_i, 1 \leq i \leq n$, are independent random variables (it is sufficient that $\varepsilon_i, 1 \leq i \leq n$, be uncorrelated) with common variance σ^2 . Then,

$$(i) \quad E(\hat{\beta}_1) = \beta_1; \quad (\hat{\beta}_1 \text{ is unbiased})$$

$$(ii) \quad E(\hat{\beta}_0) = \beta_0; \quad (\hat{\beta}_0 \text{ is unbiased})$$

$$(iii) \quad \text{Var}(\hat{\beta}_1) = \sigma^2 / S_{xx}, \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2;$$

and

$$(iv) \quad \text{Var}(\hat{\beta}_0) = \sigma^2 [1/n + \bar{x}^2 / S_{xx}].$$

(v) *In addition, if $\varepsilon_i, 1 \leq i \leq n$, are $N(0, \sigma^2)$, then*

$$\hat{\beta}_i \sim N[\beta_i, \text{Var}(\hat{\beta}_i)], \quad i = 0, 1. \quad (3.56)$$

(Note: Since the proofs of Theorem 3.1 and Theorem 3.2 are a little long, some readers may wish to skip them at this point. It is advisable, however, to know the properties stated in these theorems as they will be used repeatedly in the remainder of the Chapter.)

Proof. (i) From (3.18)

$$\begin{aligned} E(\hat{\beta}_1) &= E\left[\frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i\right] = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) E(Y_i) \\ &= \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i) \\ &= \frac{\beta_0}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{\beta_1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) x_i \\ &= 0 + \frac{\beta_1}{S_{xx}} S_{xx} = \beta_1. \end{aligned} \quad (3.57)$$

(ii) From (3.19) and (i)

$$\begin{aligned}
 E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\bar{Y}) - \bar{x}E(\hat{\beta}_1) \\
 &= E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] - \bar{x}\beta_1 = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} \\
 &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0.
 \end{aligned} \tag{3.58}$$

(iii) Again using (i) and the independence of $\{Y_i\}$, we have

$$\begin{aligned}
 Var(\hat{\beta}_1) &= Var\left[\frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i\right] \\
 &= \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i) \\
 &= \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.
 \end{aligned} \tag{3.59}$$

(iv) From (3.19) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ so that

$$\begin{aligned}
 Var(\hat{\beta}_0) &= Var(\bar{Y} - \hat{\beta}_1 \bar{x}) \\
 &= Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{Y}, \hat{\beta}_1) \\
 &= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}} - 2\bar{x} Cov(\bar{Y}, \hat{\beta}_1).
 \end{aligned}$$

Hence, to complete the proof of (iv) we need to show that $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

Now using (3.48)

$$\begin{aligned}
 Cov(\bar{Y}, \hat{\beta}_1) &= Cov\left[\bar{Y}, \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i\right] \\
 &= \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Cov(\bar{Y}, Y_i).
 \end{aligned} \tag{3.60}$$

Since, Y_i , $1 \leq i \leq n$, are independent, $Cov(Y_i, Y_j) = 0$, $i \neq j$, so that

$$\begin{aligned}
 Cov(\bar{Y}, Y_i) &= Cov\left(\frac{1}{n} \sum_{j=1}^n Y_j, Y_i\right) = \frac{1}{n} \sum_{j=1}^n Cov(Y_j, Y_i) \\
 &= \frac{1}{n} [Cov(Y_i, Y_i)] = \frac{\sigma^2}{n}
 \end{aligned} \tag{3.61}$$

and substituting (3.61) into (3.60) gives

$$Cov(\bar{Y}, \hat{\beta}_1) = \frac{\sigma^2}{n S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) = 0, \tag{3.62}$$

so that

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] \quad (3.63)$$

as required.

(v) Using (3.18) and (3.19) one easily finds that $\hat{\beta}_0$ and $\hat{\beta}_1$ is each a linear combination of the independent normal random variables Y_i . From Section 2.8 this sum is also normal with the means and variances given by (2.101)-(2.102). ■

Theorem 3.2 *Under the assumptions of Theorem 3.1, $E(s^2) = \sigma^2$. That is, s^2 is an unbiased estimator of σ^2 .*

Proof. Since $E(\hat{Y}_i) = E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = E(Y_i)$,

$$E\left[(Y_i - \hat{Y}_i)^2\right] = \text{Var}(Y_i - \hat{Y}_i). \quad (3.64)$$

Thus,

$$\begin{aligned} E\left[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2\right] &= \sum_{i=1}^n \text{Var}(Y_i - \hat{Y}_i) \\ &= \sum_{i=1}^n \text{Var}(Y_i) + \sum_{i=1}^n \text{Var}(\hat{Y}_i) - 2 \sum_{i=1}^n \text{Cov}(Y_i, \hat{Y}_i) \\ &= n\sigma^2 + \sum_{i=1}^n \text{Var}(\hat{Y}_i) - 2 \sum_{i=1}^n \text{Cov}(Y_i, \hat{Y}_i). \end{aligned} \quad (3.65)$$

We now consider $\text{Var}(\hat{Y}_i)$ which is given by

$$\begin{aligned} \text{Var}(\hat{Y}_i) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \text{Var}(\hat{\beta}_0) + x_i^2 \text{Var}(\hat{\beta}_1) + 2x_i \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] + \frac{x_i^2 \sigma^2}{S_{xx}} + 2x_i \text{Cov}(\hat{\beta}_0, \hat{\beta}_1). \end{aligned} \quad (3.66)$$

But,

$$\begin{aligned} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \text{Cov}(\bar{Y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\ &= \text{Cov}(\bar{Y}, \hat{\beta}_1) - \bar{x} \text{Cov}(\hat{\beta}_1, \hat{\beta}_1) \\ &= \text{Cov}(\bar{Y}, \hat{\beta}_1) - \bar{x} \text{Var}(\hat{\beta}_1) \end{aligned} \quad (3.67)$$

and using (3.62) and (iii) of Theorem 3.1 we get

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}. \quad (3.68)$$

Hence,

$$\begin{aligned} Var(\hat{Y}_i) &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x_i^2}{S_{xx}} - \frac{2x_i\bar{x}}{S_{xx}} \right] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]. \end{aligned} \quad (3.69)$$

Thus, the second term in (3.69) becomes

$$\begin{aligned} \sum_{i=1}^n Var(\hat{Y}_i) &= \sigma^2 + \frac{\sigma^2}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sigma^2 + \frac{\sigma^2}{S_{xx}} S_{xx} = 2\sigma^2. \end{aligned} \quad (3.70)$$

Finally, we calculate $\sum_{i=1}^n Cov(Y_i, \hat{Y}_i)$. For this we have

$$\begin{aligned} Cov(Y_i, \hat{Y}_i) &= Cov(Y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= Cov(Y_i, \hat{\beta}_0) + x_i Cov(Y_i, \hat{\beta}_1). \end{aligned} \quad (3.71)$$

But,

$$\begin{aligned} Cov(Y_i, \hat{\beta}_1) &= \frac{1}{S_{xx}} \sum_{j=1}^n (x_j - \bar{x}) Cov(Y_i, Y_j) \\ &= \frac{\sigma^2 (x_i - \bar{x})}{S_{xx}} \end{aligned} \quad (3.72)$$

and

$$\begin{aligned} Cov(Y_i, \hat{\beta}_0) &= Cov(Y_i, \bar{Y} - \bar{x}\hat{\beta}_1) \\ &= Cov(Y_i, \bar{Y}) - \bar{x}Cov(Y_i, \hat{\beta}_1) \\ &= \frac{\sigma^2}{n} - \frac{\bar{x}\sigma^2 (x_i - \bar{x})}{S_{xx}} \end{aligned} \quad (3.73)$$

so that

$$\begin{aligned} Cov(Y_i, \hat{Y}_i) &= \frac{\sigma^2}{n} - \frac{\bar{x}\sigma^2 (x_i - \bar{x})}{S_{xx}} + \frac{\sigma^2 x_i (x_i - \bar{x})}{S_{xx}} \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2 (x_i - \bar{x})^2}{S_{xx}}. \end{aligned} \quad (3.74)$$

Thus,

$$\begin{aligned} \sum_{j=1}^n Cov(Y_i, \hat{Y}_i) &= \sigma^2 + \frac{\sigma^2}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sigma^2 + \sigma^2 = 2\sigma^2. \end{aligned} \quad (3.75)$$

Using (3.75) in (3.69) gives

$$E \left[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \right] = n\sigma^2 + 2\sigma^2 - 4\sigma^2 = (n-2)\sigma^2 \quad (3.76)$$

so that

$$\begin{aligned} E(s^2) &= E \left[\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \right] \\ &= \frac{\sigma^2}{n-2} (n-2) = \sigma^2. \end{aligned} \quad (3.77)$$

■

Last we observe that if the errors $\{\varepsilon_i\}$ are independent and $N(0, \sigma^2)$, then the distribution of $(n-2)s^2/\sigma^2$ is χ^2 with $n-2$ degrees of freedom, and surprisingly, s^2 is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$. These results are important, as we shall see, in developing test procedures for the simple linear regression model. The proofs of these latter two properties are somewhat technical, and will be obtained as a consequence of a more general argument for the multiple regression model in Chapter 5. Again these facts should be remembered, even if the proofs are omitted.

3.4.1 Standard Errors of the Coefficients

Once we have estimated σ^2 , then we can obtain point estimates of the variances and standard deviations of $(\hat{\beta}_0, \hat{\beta}_1)$. From Theorem 3.1 we find that if the errors are uncorrelated and have common variance σ^2 , then $\text{Var}(\hat{\beta}_0) = \sigma^2 (\hat{\beta}_0) = \sigma^2 [1/n + \bar{x}^2/S_{xx}]$ and $\text{Var}(\hat{\beta}_1) = \sigma^2 (\hat{\beta}_1) = \sigma^2/S_{xx}$. Thus, unbiased estimates of these quantities are given by

$$\hat{\sigma}^2(\hat{\beta}_0) = s^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] = s^2 \delta_0, \quad (3.78)$$

and

$$\hat{\sigma}^2(\hat{\beta}_1) = \frac{s^2}{S_{xx}} = s^2 \delta_1, \quad (3.79)$$

where δ_0 and δ_1 are referred to as the *variance multiplication factors* (VMF). The standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ are then estimated by

$$\hat{\sigma}(\hat{\beta}_0) = s\sqrt{\delta_0}, \quad (3.80)$$

and

$$\hat{\sigma}(\hat{\beta}_1) = s\sqrt{\delta_1}. \quad (3.81)$$

$\hat{\sigma}(\hat{\beta}_0)$ and $\hat{\sigma}(\hat{\beta}_1)$ are called the *standard errors of the estimates* $(\hat{\beta}_0, \hat{\beta}_1)$. As we shall see, these standard errors play a fundamental role in obtaining confidence intervals for β_0 and β_1 and in developing hypothesis tests for these quantities.

Example 3.9 Find the standard errors of $(\hat{\beta}_0, \hat{\beta}_1)$ in Example 3.2.

Here the calculations are simple enough to be done by hand. We find from the calculations in that example that $\bar{x} = 3$ and $S_{xx} = 10$, so that $s = 0.6056$

$$\delta_0 = 1/5 + 9/10 = 11/10 \text{ and } \delta_1 = 1/10.$$

Thus,

$$\hat{\sigma}(\hat{\beta}_0) = 0.6055\sqrt{11/10} = 0.6350,$$

and

$$\hat{\sigma}(\hat{\beta}_1) = 0.6055\sqrt{1/10} = 0.1915.$$

Example 3.10 Find the standard errors of the coefficients in the drink delivery model in Example 3.4.

Here the calculations are too tedious to do by hand so we read the values directly from the computer output. They are

$$\hat{\sigma}(\hat{\beta}_0) = 1.371 \text{ and } \hat{\sigma}(\hat{\beta}_1) = 0.124.$$

Notice in this case that the standard error for $\hat{\beta}_1$ is quite small in comparison to that for $\hat{\beta}_0$ which is 1.371. This suggests that the “true” value of β_1 is probably quite close to its estimated value $\hat{\beta}_1$. The larger standard error for $\hat{\beta}_0$ indicates that we should have less confidence in the value of $\hat{\beta}_0$ as an estimate of β_0 . These ideas will be formalized in Section 3.6.

3.5 The Gauss-Markov Theorem

As we have seen, when the errors in the model (3.1) are normal, then maximum likelihood estimation is a justification for the use of least squares estimators of (β_0, β_1) . However, when the errors are not normal, then as shown in Section 3.2 this is generally not the case so it is important to ask what statistical justification there is for using least square estimation when the errors are not normal. In such circumstances, it can be shown that the least squares estimators are the *minimum variance unbiased linear estimators* of (β_0, β_1) . This optimality property, the *Gauss-Markov theorem*, is often invoked to justify least squares estimation, even for non-normal errors. On the other hand, we hasten to point out that there may very well be nonlinear and/or biased estimators of (β_0, β_1) which are more efficient than least squares for error distributions which depart markedly from normal. This is a currently active research area, often called *robust regression analysis* and some aspects of this subject can be found in [27, 65].

We now proceed to a statement and proof of the Gauss-Markov theorem for simple linear regression. The generalization to multiple regression will be taken up in Chapter 5.

Definition 3.1 Consider the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. A *linear* estimator $\hat{\beta}_j, j = 0, 1$, for β_j is one of the form

$$\hat{\beta}_j = \sum_{k=1}^n c_{jk} Y_k, \quad j = 0, 1, \quad (3.82)$$

where c_{jk} , $1 \leq k \leq n$ are given constants.

Note: Observe from (3.18) and (3.19) that the least squares estimators of (β_0, β_1) are linear.

Theorem 3.3 *Consider the model*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad 1 \leq i \leq n, \quad (3.83)$$

where ε_i , $1 \leq i \leq n$, are uncorrelated and have constant variance $\text{Var}(\varepsilon_i) = \sigma^2$, $1 \leq i \leq n$. Then the minimum variance unbiased linear estimator of β_j , $j = 0, 1$, is the least squares estimator.

Proof. We will consider the proof for β_1 , since the proof for β_0 is similar. Now suppose that

$$\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i. \quad (3.84)$$

Then, since ε_i , $1 \leq i \leq n$, are uncorrelated and have the same variance,

$$\text{Var}(\hat{\beta}_1) = \sum_{i=1}^n c_i^2 \text{Var}(Y_i) = \sigma^2 \left(\sum_{i=1}^n c_i^2 \right), \quad (3.85)$$

and for $\hat{\beta}_1$ to be unbiased we must have $E(\hat{\beta}_1) = \beta_1$. This gives

$$E(\hat{\beta}_1) = \sum_{i=1}^n c_i E(Y_i) = \sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i) = \beta_1. \quad (3.86)$$

Since (3.86) must be true for all β_0 and β_1 , we have, equating the coefficients of β_0 and β_1 in (3.86)

$$\sum_{i=1}^n c_i = 0 \quad \text{and} \quad \sum_{i=1}^n c_i x_i = 1. \quad (3.87)$$

Thus our problem reduces to minimizing $\sum_{i=1}^n c_i^2$ subject to (3.87). The solution to this problem may be found by using the technique of *Lagrange multipliers* [104, 27]. That is, we choose $(c_i, 1 \leq i \leq n, \lambda_1, \lambda_2)$ to be critical points of

$$Q = \sum_{i=1}^n c_i^2 - 2\lambda_1 \left(\sum_{i=1}^n c_i \right) - 2\lambda_2 \left(\sum_{i=1}^n c_i x_i - 1 \right). \quad (3.88)$$

From calculus this can be done by solving the equations

$$\frac{\partial Q}{\partial c_i} = 0, \quad 1 \leq i \leq n, \quad \frac{\partial Q}{\partial \lambda_1} = \frac{\partial Q}{\partial \lambda_2} = 0. \quad (3.89)$$

Carrying out the differentiations in (3.89) yields

$$2c_i - 2\lambda_1 - 2\lambda_2 x_i = 0, \quad 1 \leq i \leq n, \quad (3.90)$$

$$\sum_{i=1}^n c_i = 0, \quad \sum_{i=1}^n c_i x_i = 1. \quad (3.91)$$

To solve (3.90)-(3.91) add up the equations in (3.90) to give

$$\sum_{i=1}^n c_i - n\lambda_1 - \lambda_2 \sum_{i=1}^n x_i = 0. \quad (3.92)$$

Using $\sum_{i=1}^n c_i = 0$, it gives

$$n\lambda_1 + \lambda_2 \sum_{i=1}^n x_i = 0. \quad (3.93)$$

To get another equation involving (λ_1, λ_2) multiply each equation in (3.92) by x_i and sum again. We get

$$\sum_{i=1}^n c_i x_i - \lambda_1 \sum_{i=1}^n x_i - \lambda_2 \sum_{i=1}^n x_i^2 = 0. \quad (3.94)$$

Since $\sum_{i=1}^n c_i x_i = 1$, (3.94) becomes

$$\lambda_1 \sum_{i=1}^n x_i + \lambda_2 \sum_{i=1}^n x_i^2 = 1. \quad (3.95)$$

Solving (3.93) and (3.95) for (λ_1, λ_2) by using Cramer's rule [see Chapter 4] gives

$$\begin{aligned} \lambda_1 &= \det \begin{vmatrix} 0 & \sum_{i=1}^n x_i \\ 1 & \sum_{i=1}^n x_i^2 \end{vmatrix} \bigg/ \det \begin{vmatrix} \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} \\ &= - \frac{\sum_{i=1}^n x_i}{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right]} \end{aligned} \quad (3.96)$$

and

$$\begin{aligned} \lambda_2 &= \det \begin{vmatrix} \sum_{i=1}^n x_i & 0 \\ \sum_{i=1}^n x_i^2 & 1 \end{vmatrix} \bigg/ \det \begin{vmatrix} \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} \\ &= - \frac{n}{\left[n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i^2 \right]} \end{aligned} \quad (3.97)$$

where \det denotes the determinant.

Dividing the numerator and denominator in (3.96)-(3.97) by n and using the fact that $\sum_{i=1}^n x_i^2 - n\bar{x}^2 = S_{xx}$ gives

$$\lambda_1 = -\frac{\bar{x}}{S_{xx}}, \quad \lambda_2 = \frac{1}{S_{xx}} \quad \text{and} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \quad (3.98)$$

so that

$$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i, \quad (3.99)$$

which agrees with the expression (3.18) for the least squares estimator. ■

Note: We have really only established that $(\lambda_1, \lambda_2, c_i, 1 \leq i \leq n)$ is a critical point of $Var(\hat{\beta}_1)$. That $c_i, 1 \leq i \leq n$, actually provide a minimum will be established in Chapter 5.

3.6 Confidence Intervals for (β_0, β_1)

In most statistical situations it is desirable to have confidence intervals for unknown parameters rather than just point estimates. These interval estimates discussed in Chapter 2 provide, in a certain sense, error estimates for a given point estimator, and as we shall see, they play a vital role in regression analysis.

As in problems in elementary statistics, the knowledge of the distribution of the point estimates is crucial in obtaining good confidence intervals, and to do this we will make normality assumptions. However, in many circumstances these confidence intervals are used even if the errors are not normal. This can often be justified because the least squares estimators are linear in Y_i , $1 \leq i \leq n$, which enables one to appeal to the Central Limit Theorem to obtain the approximate normality of $(\hat{\beta}_0, \hat{\beta}_1)$, even if the errors are not.

When ε_i , $1 \leq i \leq n$, are independent and $N(0, \sigma^2)$, then as we stated after Theorem 3.2, $(n-2)s^2/\sigma^2$ is $\chi^2(n-2)$ and is independent of each of $(\hat{\beta}_0, \hat{\beta}_1)$. It then follows that the random variables

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}(\hat{\beta}_i)}, \quad i = 0, 1 \quad (3.100)$$

each have a T -distribution with $n-2$ degrees of freedom. This is true since

$$T_i = \frac{(\hat{\beta}_i - \beta_i)/\sigma(\hat{\beta}_i)}{s/\sigma}, \quad i = 0, 1 \quad (3.101)$$

where $\sigma(\hat{\beta}_i) = \sigma\sqrt{\delta_i}$, $\hat{\sigma}(\hat{\beta}_i) = s\sqrt{\delta_i}$, and δ_i is given by either (3.78) or (3.79). Then $(\hat{\beta}_i - \beta_i)/\sigma(\hat{\beta}_i)$ is $N(0, 1)$ and is independent of s/σ , which is the square root of a χ^2 random variable divided by its degrees of freedom. From Section 2.8 this random variable is well known to have a t -distribution.

If $t_{n-2, \alpha/2}$ is the upper $\alpha/2$ percentage point ($1 - \alpha/2$ percentile) of a t -distribution with $n-2$ degrees of freedom, then

$$P\{-t_{n-2, \alpha/2} \leq T_i \leq t_{n-2, \alpha/2}\} = 1 - \alpha, \quad i = 0, 1 \quad (3.102)$$

so that

$$P\left\{-t_{n-2, \alpha/2} \leq \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}(\hat{\beta}_i)} \leq t_{n-2, \alpha/2}\right\} = 1 - \alpha, \quad i = 0, 1. \quad (3.103)$$

Solving the inequality in (3.103) for β_i shows that

$$P\left\{\hat{\beta}_i - t_{n-2, \alpha/2}\hat{\sigma}(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{n-2, \alpha/2}\hat{\sigma}(\hat{\beta}_i)\right\} = 1 - \alpha, \quad i = 0, 1. \quad (3.104)$$

According to the definition of a confidence interval, the pair of random variables

$$(\hat{\beta}_i - t_{n-2, \alpha/2}\hat{\sigma}(\hat{\beta}_i), \hat{\beta}_i + t_{n-2, \alpha/2}\hat{\sigma}(\hat{\beta}_i)), \quad i = 0, 1 \quad (3.105)$$

is a $(1 - \alpha) \times 100\%$ confidence interval for β_i , $i = 0, 1$.

For a given sample the value of the confidence interval is obtained by substituting the values of the estimates $\hat{\beta}_0, \hat{\beta}_1$ and s into (3.105). As is customary, we will make no notational or verbal distinction between the confidence interval and its value. For the convenience of the reader the specific formulas obtained by using (3.106) and (3.107) are given below:

(1) Confidence interval for β_0 :

$$\left(\hat{\beta}_0 \pm t_{n-2, \alpha/2} \cdot s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right). \quad (3.106)$$

(2) Confidence interval for β_1 :

$$\left(\hat{\beta}_1 \pm t_{n-2, \alpha/2} \cdot \frac{s}{\sqrt{S_{xx}}} \right). \quad (3.107)$$

One should observe from (3.106) and (3.107) that in most practical cases, the confidence interval for β_0 will be wider than that for β_1 . That is, the slope is generally estimated with greater precision than the intercept. This will be apparent in the following numerical examples.

Example 3.11 Find 95% confidence intervals for (β_0, β_1) in the drug and reaction time data in Example 3.2.

From Table A.2 we find that $t_{3, .025} = 3.182$ and from (3.106) and (3.107) 95% confidence intervals are given by

$$\hat{\beta}_0 \pm 3.182 \times \hat{\sigma} \left(\hat{\beta}_0 \right) = -0.1 \pm 3.182 \times 0.6350 = (-2.1206, 1.9206)$$

and

$$\hat{\beta}_1 \pm 3.182 \times \hat{\sigma} \left(\hat{\beta}_1 \right) = 0.7 \pm 3.182 \times 0.1915 = (0.0906, 1.3094).$$

Example 3.12 Find 95% confidence intervals for (β_0, β_1) in the soft drink delivery model in Example 3.4.

Here $n = 25$, $n - 2 = 23$ and from Table A.2 $t_{23, .025} = 2.069$ so that using the results from the computer output 95% confidence intervals are given by

$$\beta_0 : (3.321 \pm 2.069 \times 1.371) = (0.4844, 6.1576)$$

and

$$\beta_1 : (2.176 \pm 2.069 \times 0.124) = (1.9190, 2.433).$$

Note that the rather narrow confidence interval for β_1 suggests that our point estimate for the slope is probably quite reliable, while that for the intercept is much less so. This is a numerical verification of the observations made above concerning the sizes of the *variance multiplication factors* δ_0 and δ_1 .

One can also obtain confidence intervals for σ . However, such confidence intervals appear to be rarely used in practice. The details are left to the reader.

3.6.1 Simultaneous Confidence Intervals

Occasionally it is helpful to obtain joint confidence intervals (or more generally a joint confidence region) for (β_0, β_1) , because (3.106) and (3.107) will not hold simultaneously with the same level of confidence.

A number of procedures have been devised to solve this problem and we shall consider two: rectangular regions obtained through the use of Bonferroni's inequality and exact confidence ellipses. Other methods may be found in [101, 47]. We begin our discussion with a formal definition of a joint confidence region for p parameters.

Definition 3.2 Let $(\theta_1, \theta_2, \dots, \theta_p)$ be p unknown parameters of a given distribution. A joint confidence region for $(\theta_1, \theta_2, \dots, \theta_p)$, with confidence at least $(1 - \alpha) \times 100\%$ is a random region \mathcal{C} in \mathbb{R}^p (the set of all real p -tuples) not depending on $(\theta_1, \theta_2, \dots, \theta_p)$ such that

$$P\{(\theta_1, \theta_2, \dots, \theta_p) \in \mathcal{C}\} \geq 1 - \alpha, \quad (3.108)$$

for all possible values of $(\theta_1, \theta_2, \dots, \theta_p)$. If (3.108) holds with equality, then \mathcal{C} will be called an *exact* $(1 - \alpha) \times 100\%$ joint confidence region for $(\theta_1, \theta_2, \dots, \theta_p)$.

Note: If $p = 1$ and \mathcal{C} is the region determined by the pair of estimators $(\hat{\theta}_L, \hat{\theta}_U)$, $\hat{\theta}_L \leq \hat{\theta}_U$ then the confidence region determined by $(\hat{\theta}_L, \hat{\theta}_U)$ is just the usual notion of a confidence interval for θ .

One simple method for generating joint confidence regions is to take rectangular regions of the form

$$\mathcal{C} = \prod_{i=1}^p (\hat{\theta}_{L_i}, \hat{\theta}_{U_i}) \quad (3.109)$$

where $(\hat{\theta}_{L_i}, \hat{\theta}_{U_i})$ is $(1 - \alpha/p) \times 100\%$ confidence interval for θ_i .

The validity of (3.109) follows from a well known inequality in probability theory, the first Bonferroni inequality. We will state and prove this inequality in Theorem 3.4 and then use it to establish (3.109).

Theorem 3.4 (Bonferroni's inequality) *Let E_1, E_2, \dots, E_p be p events in a probability space Ω with $P\{E_i\} = 1 - \alpha_i, 1 \leq i \leq p$, then*

$$P\left\{\bigcap_{i=1}^p E_i\right\} \geq 1 - \sum_{i=1}^p \alpha_i. \quad (3.110)$$

Proof. From de Morgan's law

$$P\left\{\bigcap_{i=1}^p E_i\right\} = P\left\{\overline{\bigcup_{i=1}^p \overline{E_i}}\right\} = 1 - P\left\{\bigcup_{i=1}^p \overline{E_i}\right\}. \quad (3.111)$$

Thus,

$$P\left\{\bigcup_{i=1}^p \overline{E_i}\right\} = 1 - P\left\{\bigcap_{i=1}^p E_i\right\}. \quad (3.112)$$

Now it is well known that

$$P \left\{ \bigcup_{i=1}^p \overline{E}_i \right\} \leq \sum_{i=1}^p P(\overline{E}_i) = \sum_{i=1}^p [1 - P(E_i)] = \sum_{i=1}^p \alpha_i. \quad (3.113)$$

From (3.112) and (3.113) we get

$$1 - P \left\{ \bigcap_{i=1}^p E_i \right\} \leq \sum_{i=1}^p \alpha_i \quad (3.114)$$

so that

$$P \left\{ \bigcap_{i=1}^p E_i \right\} \geq 1 - \sum_{i=1}^p \alpha_i, \quad (3.115)$$

as required. ■

We now use (3.110) to establish (3.109). For this let $(\hat{\theta}_{L_i}, \hat{\theta}_{U_i})$ be a $(1 - \alpha/p) \times 100\%$ confidence interval for θ_i , $1 \leq i \leq p$. Then

$$P \left\{ \hat{\theta}_{L_i} \leq \theta_i \leq \hat{\theta}_{U_i} \right\} = 1 - \alpha/p. \quad (3.116)$$

Let $E_i = \left\{ \hat{\theta}_{L_i} \leq \theta_i \leq \hat{\theta}_{U_i} \right\}$ so that

$$P \left\{ \prod_{i=1}^p \left(\hat{\theta}_{L_i}, \hat{\theta}_{U_i} \right) \right\} = P \left\{ \bigcap_{i=1}^p E_i \right\}. \quad (3.117)$$

From (3.116) and (3.117)

$$P \left\{ \bigcap_{i=1}^p E_i \right\} \geq 1 - p \cdot \frac{\alpha}{p} = 1 - \alpha, \quad (3.118)$$

so that \mathcal{C} is a joint confidence region with at least $(1 - \alpha) \times 100\%$ confidence since

$$\bigcap_{i=1}^p E_i = \{(\theta_1, \theta_2, \dots, \theta_p) \in \mathcal{C}\}. \quad (3.119)$$

Specializing this argument to the case of the simple linear regression model we know that

$$\left\{ \beta_i \pm t_{n-2, \alpha/4} \hat{\sigma} \left(\hat{\beta}_i \right) \right\}, \quad i = 0, 1 \quad (3.120)$$

is a $1 - \alpha/2$ confidence interval for β_i , $i = 0, 1$. Thus the rectangle determined by

$$\left\{ \hat{\beta}_0 \pm t_{n-2, \alpha/4} \hat{\sigma} \left(\hat{\beta}_0 \right) \right\} \times \left\{ \hat{\beta}_1 \pm t_{n-2, \alpha/4} \hat{\sigma} \left(\hat{\beta}_1 \right) \right\} \quad (3.121)$$

is a joint confidence region for (β_0, β_1) with at least $(1 - \alpha) \times 100\%$ confidence.

If one uses more sophisticated arguments, then it is possible to obtain “smaller” confidence regions than those determined by (3.121). For (β_0, β_1) these regions are ellipses rather than rectangles.

Theorem 3.5 *For the simple linear regression model with independent $N(0, \sigma^2)$ errors, a joint $(1 - \alpha) \times 100\%$ confidence region for (β_0, β_1) is given by the set*

$$n(\hat{\beta}_0 - \beta_0)^2 + 2 \sum_{i=1}^n x_i (\hat{\beta}_0 - \beta_0) (\hat{\beta}_1 - \beta_1) + \sum_{i=1}^n x_i^2 (\hat{\beta}_1 - \beta_1)^2 \leq 2s^2 f_{\alpha, 2, n-2} \quad (3.122)$$

where $f_{\alpha, 2, n-2} = P\{F_{2, n-2} \geq \alpha\}$ and $F_{2, n-2}$ is an F random variable with $(2, n-2)$ degrees of freedom.

Proof. We will also obtain this as a particular case of a more general result for multiple regression models in Chapter 5. ■

3.7 Hypothesis Tests for (β_0, β_1)

So far we have assumed that the model that relates y to x is of the form in (3.1) and all subsequent results have been based on this assumption. We now turn our attention to assessing the validity of these assumptions.

Assuming for the time being that the model is linear and that x is the only possible explanatory variable we consider the question of determining whether x is helpful in explaining the variation in y . That is, we wish to distinguish between the models

$$Y_i = \beta_0 + \varepsilon_i \quad (3.123)$$

and

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (3.124)$$

In parametric terms we want to test

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0. \quad (3.125)$$

In accepting H_0 , we can conclude that x is of no use in explaining the variation of y , whereas accepting H_1 suggests that it is. However keep in mind that one can only come to these conclusions if one truly knows that the model is linear. In practical situations one usually does not know this for certain, and a more cautious interpretation is called for. In particular, accepting H_0 may not mean that x is not a useful explanatory variable. It may only indicate that the variation of y with x may not contain a linear component. On the other hand, accepting H_1 suggests that there is a linear trend of y with x but other types of variation may be present as well.

To develop a test of H_0 against H_1 we use the confidence intervals developed for β_1 in the previous section. These confidence intervals, depending on the confidence level, may be viewed as giving a plausible range of values for the true slope β_1 . If zero is in one of these intervals, then at the appropriate level of confidence zero is a plausible value of β_1 and we accept H_0 . If zero is not in the interval, then zero is not a plausible value β_1 and we reject H_0 . Thus we reject H_0 if

$$\hat{\beta}_1 + t_{n-2, \alpha/2} \hat{\sigma} (\hat{\beta}_1) < 0 \text{ or } \hat{\beta}_1 - t_{n-2, \alpha/2} \hat{\sigma} (\hat{\beta}_1) > 0. \quad (3.126)$$

Rearranging we find that the critical region for rejecting H_0 is given by

$$\frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} < -t_{n-2, \alpha/2} \text{ or } \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} > t_{n-2, \alpha/2} \quad (3.127)$$

which is equivalent to

$$\left| \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \right| > t_{n-2, \alpha/2}, \quad (3.128)$$

where $\hat{\sigma}(\hat{\beta}_1) = s/\sqrt{S_{xx}}$. The quantity $\hat{\beta}_1/\hat{\sigma}(\hat{\beta}_1)$ is referred to as the *observed t value*, so our test is to reject H_0 if $|t \text{ (observed)}| > t \text{ (tabulated)}$. From our discussion in Section 2.13 we see that this test has significance level α .

Example 3.13 For the soft drink delivery data in Example 3.4 can we reject H_0 at the 5% level of significance?

From the computer output we find that the observed t value for β_1 is 17.55. Since $t_{23, 0.025} = 2.069$ we can reject H_0 at the 5% level of significance. In fact, the probability that $|t| > 17.55$ if H_0 is true is less than 0.0001 so that we can accept H_1 with at least 99.99% confidence.

An intuitive interpretation of the t -test for β_1 can be given in the following way. Note that $\hat{\sigma}(\hat{\beta}_1)$ estimates the error in estimating β_1 , so that $t = \left| \hat{\beta}_1 / \hat{\sigma}(\hat{\beta}_1) \right|$ represents the reciprocal of a “relative” error in estimating β_1 . If β_1 is well estimated, then we expect $\hat{\sigma}(\hat{\beta}_1)$ to be “small” in relation to $\hat{\beta}_1$, hence t is “large”. Thus, the t -test says that we should accept $\beta_1 \neq 0$ provided that the β_1 is well estimated relative to its error.

Most modern regression programs print out the p -value $P\{T > |t \text{ (observed)}|\}$ where T is a T random variable with $n - 2$ degrees of freedom. If this information is not available, a quick “eyeball” test can be given. Note from Table A.3 that $t_{25, 0.025} = 2.060$ and this value does not change significantly as the degrees of freedom increases. This is a consequence of the fact, observed in Chapter 2, that the limit of a t density as the number of degrees of freedom increases is a $N(0, 1)$ density with $z_{0.025} = 1.96$.

This leads to a simple rule for rejecting H_0 for a sample size $n > 20$. Reject H_0 if $|t \text{ (observed)}| > 2$. This rule will be extended to multiple regression in Chapter 5.

In situations where one has a pre-conceived notion as to the true value b_1 of β_1 , the t -test for (3.125) can be extended to test

$$H_0 : \beta_1 = b_1 \text{ against } H_1 : \beta_1 \neq b_1 \quad (3.129)$$

by forming the t statistic

$$t \text{ (observed)} = \frac{\hat{\beta}_1 - b_1}{\hat{\sigma}(\hat{\beta}_1 - b_1)} \quad (3.130)$$

and rejecting H_0 if $|t \text{ (observed)}| > t_{n-2, \alpha/2}$ for a test with significance level α . Since $\hat{\sigma}(\hat{\beta}_1 - b_1) = \hat{\sigma}(\hat{\beta}_1)$ because, $Var(\hat{\beta}_1 - b_1) = Var(\hat{\beta}_1)$,

$$t \text{ (observed)} = \frac{\hat{\beta}_1 - b_1}{\hat{\sigma}(\hat{\beta}_1)}. \quad (3.131)$$

As for (3.129), this test can be derived by a confidence interval argument. We leave the details to the reader.

As for testing for the value of the slope of $y = \beta_0 + \beta_1 x + \varepsilon$ one commonly tests for the significance of the intercept. Here, the situation of most interest is determining where the true line passes through the origin $(0, 0)$. This leads to the test

$$H_0 : \beta_0 = 0 \text{ against } H_1 : \beta_0 \neq 0. \quad (3.132)$$

Accepting H_0 suggests that the simpler model $y = \beta_1 x + \varepsilon$ explains the data better than $y = \beta_0 + \beta_1 x + \varepsilon$. If H_0 is true, the data may be refitted by this simpler model. We take this matter up in greater detail in Section 3.8.

As for β_0 , we can test (3.132) by using the t statistic

$$t \text{ (observed)} = \frac{\hat{\beta}_0}{\hat{\sigma}(\hat{\beta}_0)} \quad (3.133)$$

and reject H_0 at level α if $|t \text{ (observed)}| > t_{n-2, \alpha/2}$. As for β_1 an “eyeball” test is to reject H_0 provided that

$$|t \text{ (observed)}| > 2. \quad (3.134)$$

Again, this test can be derived by using a confidence interval argument. The details are left to the reader.

One sided-tests for the regression coefficients can be obtained by using one-sided confidence intervals as indicated in Section 2.13.

3.8 The ANOVA Approach to Testing

In the previous section we have seen how a t -test could be used to test for the significance of the regression. Here we will develop an equivalent F -test, which, as we shall see, can be generalized to test for the overall significance of the regression in the multiple regression case as well. This will be useful since the corresponding tests for the significance of the individual coefficients may be misleading.

The idea behind this method, called the *analysis of variance* (ANOVA) approach to testing, is to determine how much of the variability in the observations (y_1, y_2, \dots, y_n) is “explained” by the regression line. For this purpose a natural measure of variability in the data is the quantity, the *total sum of squares*;

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3.135)$$

which is $n - 1$ times the sample variance of (y_1, y_2, \dots, y_n) . If the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ fits the data well, then $\hat{y}_i \simeq y_i$ so that

$$\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \simeq \sum_{i=1}^n (y_i - \bar{y})^2. \quad (3.136)$$

Since $\bar{\hat{y}} = \bar{y}$ if $\hat{\beta}_0 \neq 0$ (this will be shown in Theorem 3.8), then

$$\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (3.137)$$

so that the ratio

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.138)$$

measures of the fraction of the variability in (y_1, y_2, \dots, y_n) that can be accounted for by the regression line. For a good fit we would like this quantity, usually called the *coefficient of determination*, to be close to one. Since R^2 can also be shown to be the square of the *sample correlation coefficient* between (y_1, y_2, \dots, y_n) and $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$, this leads further credibility to R^2 as a measure of “goodness of fit.”

Because R^2 is a function of the sample values (y_1, y_2, \dots, y_n) , it is a random variable, so if we know its distribution it would seem plausible to test for the significance of the regression by rejecting $H_0 : \beta_1 = 0$, if R^2 is sufficiently close to one. Since any monotone function of R^2 gives an equivalent test, it is convenient to consider

$$F = \frac{(n-2) R^2}{1 - R^2}. \quad (3.139)$$

The reason for this rather arbitrary choice will become apparent after we develop some of the basic properties of R^2 and F . These are given in Theorem 3.6

Theorem 3.6 *Assume that $SST \neq 0$. Then*

- (i) $0 \leq R^2 \leq 1$;
- (ii) $R^2 = 1 - SSE/SST$;
- (iii) $R^2 = 1$ if and only if $\hat{y}_i = y_i$, $1 \leq i \leq n$ (i.e., the data points all lie on the regression line).
- (iv) If $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$, then $R^2 = \rho^2(\mathbf{x}, \mathbf{y})$, where

$$\rho^2(\mathbf{x}, \mathbf{y}) = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{S_{xx}S_{yy}} \quad (3.140)$$

is the square of the sample correlation coefficient between (\mathbf{x}, \mathbf{y}) .

- (v) $F = SSR/s^2 = T^2$
- (vi) If the errors ε_i , $1 \leq i \leq n$, are independent $N(0, \sigma^2)$ random variables and $\beta_1 = 0$ in (3.1), then F has an F -distribution with $(1, n-2)$ degrees of freedom.

Note that it is the relation in (v) that reveals our interest in R^2 and F . From (v) and (vi) we see that using any of T , R^2 or F provides us with equivalent statistics for testing the significance of the regression

We also point out that the size of R^2 is almost always examined when doing a least squares fit. It provides a rough “eyeball” goodness of fit test; i.e., the closer R^2 is to one, the “better” the regression line fits the data. (However, some caution is needed in this respect and we will return to this point later.) The F ratio may then be viewed as a statistic for testing the significance of large values of R^2 .

Before proving Theorem 3.6 we point out that many of the details depend on establishing the following decomposition of SST:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.141)$$

Equation (3.141) is usually written in the form

$$SST = SSR + SSE \quad (3.142)$$

where SST is called the *total sum of squares* (also the total adjusted-for the mean sum of square), SSR is the *regression sum of squares* and of course SSE is the *residual sum of squares*.

To establish (3.141) we need a lemma concerning the residuals from the least squares fit.

Lemma 3.1 *Let $\hat{\varepsilon}_i = y_i - \hat{y}_i$, $1 \leq i \leq n$, denote the residuals from the least squares fit $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $1 \leq i \leq n$, to (y_1, y_2, \dots, y_n) . Then,*

$$(i) \sum_{i=1}^n \hat{\varepsilon}_i = 0;$$

$$(ii) \bar{\hat{y}} = \bar{y};$$

$$(iii) \sum_{i=1}^n \hat{\varepsilon}_i \hat{y}_i = 0.$$

Proof. (i) From the first normal equation $\partial S / \partial \beta_0 = 0$, where S is given by (3.9), we get

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n \hat{\varepsilon}_i = 0. \quad (3.143)$$

(ii) From (i) $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$, so dividing both of these sums by n gives (ii).

(iii) From the second normal equation $\partial S / \partial \beta_1 = 0$ we get

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^n x_i \hat{\varepsilon}_i = 0. \quad (3.144)$$

Thus,

$$\begin{aligned} \sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i &= \sum_{i=1}^n \hat{\varepsilon}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum_{i=1}^n \hat{\varepsilon}_i \hat{\beta}_0 + \sum_{i=1}^n x_i \hat{\varepsilon}_i \hat{\beta}_1 \\ &= \hat{\beta}_0 \sum_{i=1}^n \hat{\varepsilon}_i + \hat{\beta}_1 \sum_{i=1}^n x_i \hat{\varepsilon}_i = 0. \end{aligned} \quad (3.145)$$

■

Note that the fact that $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ provides a convenient check on the accuracy of least squares calculations. A rough “rule of thumb” is that one can expect that the number of valid significant figures to the right of the decimal point in $\hat{\beta}_0, \hat{\beta}_1$, etc.

is about the same as the number of zeros to the right of the decimal point in the sum of the residuals.

We now use Lemma 3.1 to prove Equation (3.142). For this we write

$$\begin{aligned}
 \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
 &= SSE + SSR + 2 \sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}). \tag{3.146}
 \end{aligned}$$

Now from Lemma 3.1 we get

$$\sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}) = \sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i - \bar{y} \sum_{i=1}^n \hat{\varepsilon}_i = 0 + 0 = 0. \tag{3.147}$$

Thus,

$$SST = SSR + SSE.$$

Proof of Theorem 3.6. (i) From (3.142)

$$0 \leq R^2 = \frac{SSR}{SST} \leq \frac{SST}{SST} = 1. \tag{3.148}$$

(ii) From (3.142) $SSR = SST - SSE$ so that

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}. \tag{3.149}$$

(iii) From (ii) $R^2 = 1$ if and only if $SSE = 0$. Since $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 0$, then $y_i = \hat{y}_i$, $i = 1, 2, \dots, n$.

(iv) From the definition of \hat{y}_i we get

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 (x_i - \bar{x}), \tag{3.150}$$

so that

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}. \tag{3.151}$$

Using the definition of the sample correlation coefficient

$$\rho^2(\mathbf{x}, \mathbf{y}) = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{S_{xx} S_{yy}} \tag{3.152}$$

and the formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}, \tag{3.153}$$

$$\rho^2(\mathbf{x}, \mathbf{y}) = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{SSR}{SST} = R^2, \tag{3.154}$$

since $S_{yy} = \sum_{i=1}^n (y_i - \bar{y}) = SST$.

(v) From the definition of F ,

$$F = \frac{(n-2)R^2}{1-R^2} = \frac{(n-2)(SSR/SST)}{(SSE/SST)} = \frac{SSR}{SSE/(n-2)} = \frac{SSR}{s^2}. \quad (3.155)$$

Also from (3.128)

$$T^2 = \frac{\hat{\beta}_1^2}{s^2/S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{s^2} = \frac{SSR}{s^2} = F. \quad (3.156)$$

(vi) If $\{\varepsilon_i\}_{i=1}^n$ are independent $N(0, \sigma^2)$, then as was stated in Section 3.2 T has a t -distribution with $n-2$ degrees of freedom. Since $F = T^2$, it follows that F has an F -distribution with $(1, n-2)$ degrees of freedom. ■

It is common practice to summarize many of the results of Theorem 3.6 in a tabular format like that in Table 3.8. This table is usually referred to as the *ANOVA* (short for analysis of variance) *table* for the regression analysis and is standard output from all of the commonly available regression packages.

Table 3.8 Analysis of Variance (ANOVA) Table				
Source	df	SS	MS	F
Regression	1	SSR	$MSR = SSR$	$\frac{MSR}{MSE}$
Residual	$n-2$	SSE	$MSE = \frac{SSE}{n-2} = s^2$	
Total	$n-1$	SST		
$R^2 = \frac{SSR}{SST}$				

The meaning of the ANOVA Table headings is as follows:

Source: source of the sum of squares.

df: the number of degrees of freedom associated with the corresponding sum of squares.

SS: Abbreviation for “sum of squares.”

MS: Abbreviation for “mean squares,” defined by SS/df .

F : The F -ratio, defined by $MSR/MSE = MSR/s^2$.

R^2 : SSR/SST . (Note that some authors express R^2 as a percentage - the percentage of variation in the observations explained by the regression line.)

Classically, the df associated with each sum of squares (SS) derives from the fact that under the standard normality assumptions about the errors, SSR/σ^2 is $\chi^2(1)$, SSE/σ^2 is $\chi^2(n-2)$ and SST/σ^2 is $\chi^2(n-1)$ when the null hypothesis $\beta_1 = 0$ is true. The degrees of freedom are then the degrees of freedom associated with the corresponding χ^2 random variables. However, even if the errors are not normal, it is still customary to associate the χ^2 degrees of freedom in the normal case with the corresponding SS in the non-normal case as well. For example, one can think that in SSE the n residuals are free to vary subject only to the conditions imposed by the least squares equations:

$\sum_{i=1}^n \hat{\varepsilon}_i = 0$ and $\sum_{i=1}^n x_i \hat{\varepsilon}_i = 0$. Then there are $n - 2$ degrees of freedom associated with SSE . Similarly, $y_i - \bar{y}$ are constrained by the fact that $\sum_{i=1}^n (y_i - \bar{y}) = 0$ so SST has $n - 1$ d.f. Since $SSR = SST - SSE$, the degrees of freedom of SSR is $(n - 2) - (n - 1) = 1$.

Further extensions of this table for simple linear regression will be given in Section 3.9 and for multiple regression in Chapter 5.

Before giving several numerical examples of ANOVA tables we prove the distribution properties of SSR/σ^2 , SSE/σ^2 and SST/σ^2 mentioned above.

Theorem 3.7 *If the errors ε_i , $1 \leq i \leq n$, in (3.1) are independent $N(0, \sigma^2)$, then if $H_0 : \beta_1 = 0$ is true, SSR/σ^2 is $\chi^2(1)$, SSE/σ^2 is $\chi^2(n - 2)$ and SST/σ^2 is $\chi^2(n - 1)$.*

Proof. As shown in the previous theorem, $SSR = \hat{\beta}_1^2 S_{xx}$ so that

$$\frac{SSR}{\sigma^2} = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \left(\frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\sigma} \right)^2. \quad (3.157)$$

From Theorem 3.1 $\hat{\beta}_1$ is $N(\beta_1, \sigma^2/S_{xx})$ so that $\hat{\beta}_1 \sqrt{S_{xx}}/\sigma$ is $N(\beta_1 \sqrt{S_{xx}}/\sigma, 1)$.

Now if $\beta_1 = 0$, then $\hat{\beta}_1 \sqrt{S_{xx}}/\sigma$ is $N(0, 1)$ and SSR/σ^2 is the square of a $N(0, 1)$ random variable and so is $\chi^2(1)$. As we pointed out in Section 3.3 SSE/σ^2 is $\chi^2(n - 2)$, regardless of whether $\beta_1 = 0$. (A proof will be given in Chapter 5.) Also, SSE/σ^2 is independent of $\hat{\beta}_1$ and so of $\hat{\beta}_1^2$. (See Section 3.6.) Thus, SSR/σ^2 and SSE/σ^2 are independent random variables. Since

$$\frac{SST}{\sigma^2} = \frac{SSR}{\sigma^2} + \frac{SSE}{\sigma^2}, \quad (3.158)$$

SST/σ^2 is the sum of two independent χ^2 random variables and so is χ^2 with $(n - 2) + 1 = n - 1$ degrees of freedom. ■

Example 3.14 Calculate the ANOVA table for the data in Example 3.2 and use it to determine the percentage of variability in the data explained by the regression line. Also test if $\beta_1 = 0$ under the assumption that the errors are independent $N(0, \sigma^2)$.

Using the calculations made in Examples 3.2 and 3.6, we find that

$$\begin{aligned} SST &= \sum_{i=1}^5 (y_i - \bar{y})^2 = 6 \\ SSR &= \hat{\beta}_1^2 S_{xx} = (0.7)^2 \times 10 = 4.9 \end{aligned}$$

and

$$SSE = SST - SSR = 6 - 4.9 = 1.1.$$

Thus,

$$R^2 = \frac{SSR}{SST} = \frac{4.9}{6} = 0.8167$$

and

$$F = \frac{SSR}{s^2} = \frac{4.9}{0.3667} = 13.37.$$

Arranging these data in the ANOVA table gives:

Table 3.9 ANOVA Table for Drug Data				
Source	df	Sum of Squares	Mean Squares	F
Regression	1	4.9	4.9	13.37
Residual	3	1.1	0.3667	
Total	4	6.0		
$R^2 = 0.8167$				

From the ANOVA table we see that since $R^2 = 0.8167$, the regression line explains 81.67% of the variability in the data.

To test for the significance of the regression we use $\alpha = 0.05$. If the computed F value 13.37 exceeds the tabulated $f_{0.05,1,3}$ value then we can reject H_0 at this level, otherwise H_0 is accepted. From Table A.4 $f_{0.05,1,3} = 10.13$, so that H_0 is rejected and the regression appears to be significant at this level.

Note that as the general theory shows $F = T^2$, and so using F or T , gives equivalent tests in this instance.

Example 3.15 Compute the ANOVA table for the drink delivery data in Example 3.4 and use it to draw conclusions about the validity of this model.

Again, due to the tediousness of the calculations, we resort to computer output for the necessary calculations. This yields Table 3.10.

Table 3.10 ANOVA Table for Delivery Data				
Source	df	Sum of Squares	Mean Squares	F
Regression	1	5,382,409	5,382,409	307.8
Residual	23	402,134	17,484	
Total	24	5,784,543		
$R^2 = 0.9305$				

We note first that the R^2 value shows that the model explains over 93% of the variability in the data. This indicates that the fit is quite good and the large observed F value 307.8 is significant at the 0.0001 level. This, coupled with the small value of $s = 4.18$, compared to the mean value of $y = 22.38$ suggests that this model appears to explain virtually all of the variability in the data and could in its present form probably make a useful predictor. We will return to this point later.

Example 3.16 Construct the ANOVA table for the tractor maintenance data in Example 3.5.

Here, because hand calculations are inconvenient, we use the data in the computer output given for that example.

Table 3.11 ANOVA Table for Tractor Data				
Source	df	Sum of Squares	Mean Squares	F
Regression	1	1,099,635	1,099,635	13.68
Residual	15	1,205,407	80,360	
Total	16	2,305,042		
$R^2 = 0.477$				

Note that even though R^2 is relatively small, that $F = 13.68 > f_{0.05,1,15} = 4.54$ so that the regression is significant at the 5% level. As is indicated in Exercise 3.8 when there are repeat values of the observations at a given x value, R^2 will generally always be smaller than for a similar data set with no repeat values. Again this suggests that the analyst should be cautious in using only R^2 as a measure of fit. Low R^2 values do not necessarily mean that the regression is not significant. There may be just a large amount of unexplainable random variation in the data.

Example 3.17 Even though large values of R^2 indicate an approximate linear relation between x and y one must be cautious in interpreting a high degree of correlation as implying causation. An interesting example concerns data collected by Kendall and Yule [70]. From 1924 to 1937 the number of certified mental defectives/100,000 population in England was recorded along with the number of radio licenses. The data are shown in Table 3.12.

Table 3.12 Radio Licenses and Mental Defects Data		
Year	Radios in millions (x)	Mental defectives (y)
1924	1.350	8
1925	1.960	8
1926	2.270	9
1927	2.483	10
1928	2.730	11
1929	3.091	11
1930	3.647	12
1931	4.620	16
1932	5.497	18
1933	6.260	19
1934	7.012	20
1935	7.618	21
1936	8.131	22
1937	8.593	23

A scatter plot shown in Figure 3.10 shows a clear linear trend in the number of mental defectives as the number of radios increases.

Based on this, the data were fit by least squares with the following results;

$$\begin{aligned}\hat{\beta}_0 &= 4.5822, t_0 = 10.82, p < 0.000, \\ \hat{\beta}_1 &= 2.20417, t_1 = 27.31, p < 0.000\end{aligned}$$

and the ANOVA table is given in Table 3.13.

Table 3.13 ANOVA Table for Radio Licenses Data				
Source	df	Sum of Squares	Mean Squares	F
Regression	1	393.39	393.39	745.96
Residual	12	6.33	0.53	
Total	13	399.72		
$R^2 = 0.984$				

These results shows a highly significant linear relation between x and y with the least squares line explaining 98.4% of the variation in the data. On the basis of these figures

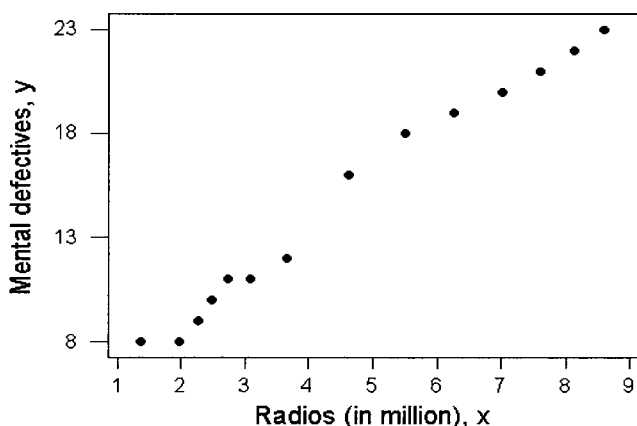


Figure 3.10: Scatter plot for radio and mental defectives data

one might conclude that radios were making people crazy. Although that is certainly possible, a more benign explanation can be given. That is, if both x and y were each increasing linearly with time, then clearly y would increase linearly with x . This latter explanation seems more plausible, since radios became more available with time and better diagnostic procedures enabled one to do a better job of identifying people with mental problems.

Since the value of R^2 is so frequently used as a measure of goodness fit, it is important to caution the reader that sometimes large values of R^2 may occur merely because the range of the data is large, producing large values of S_{xx} . To see this, we note that a result of Hahn [48] shows that

$$E(R^2) \simeq \frac{\beta_1^2 S_{xx}}{\beta_1^2 S_{xx} + \sigma^2} \quad (3.159)$$

and this is easily shown to be an increasing function of S_{xx} . Thus, a large spread in the independent variable (x_1, x_2, \dots, x_n) may produce a large value of R^2 having little to do with the quality of fit. Also note that $E(R^2)$ is an increasing function of β_1^2 . Thus, models with “large” slopes will generally produce larger values of R^2 than those with smaller ones. Again, we observe that a large value of R^2 may occur which is not indicative of the quality of the fit.

In general, we need to use multiple criteria in assessing the quality of the fit. Among these criteria are:

- (i) ‘large’ R^2 ,
- (ii) ‘large’ F or $|t|$ values,
- (iii) ‘small’ values of s^2 relative to \bar{y} , the mean of the observations.

As we shall see, other criteria need to be examined as well, and these will be discussed as we proceed.

3.8.1 Regression Through the Origin

So far we have considered the simple linear regressions model in the form of (3.1) where initially it is assumed that $\beta_0 \neq 0$. However, there are some circumstances where the appropriate model requires that $\beta_0 = 0$. In this case we have a situation which differs slightly from then when $\beta_0 \neq 0$ and we will discuss the similarities and differences in estimation and inference in this section. As noted previously, the regression problem with $\beta_0 = 0$ is usually called *regression through the origin*.

First, where would the model

$$Y_i = \beta_1 x_i + \varepsilon_i, \quad 1 \leq i \leq n, \quad (3.160)$$

be more appropriate than (3.1). There are two such circumstances.

- (i) If it is known a priori from physical considerations that $E(Y_0) = \beta_0 = 0$, then there is no point in using a degree of freedom to estimate β_0 , since this will generally decrease the accuracy in estimating σ^2 and so the accuracy in estimating β_1 , will generally be decreased as well.

(This is, as we shall see in Chapter 8, a general property of including extraneous variables in the regression model.)

As a hypothetical example, suppose we wished to make a model of the weight y of a given population as a function of weight x . If we knew that the model was *linear* at $x = 0$, then certainly a person of zero height would have zero weight and choosing $\beta_0 = 0$ is appropriate.

- (ii) If we believe initially that $\beta_0 \neq 0$ and a t -test suggests on the basis of the observed data that $\beta_0 = 0$, then β_0 might be eliminated from the model. Since in many practical cases one cannot be sure that the model is valid near the origin, some statisticians insist that the intercept always be kept in even if it appears statistically insignificant from zero.

One should be cautioned, particularly in using (3.1) that choosing the intercept $\beta_0 = 0$ initially, may not be correct even if physically $E(Y_0) = 0$. Unless we know for certain that the model is linear near $x = 0$, setting $\beta_0 = 0$ may lead to badly biased estimates of β_1 if the independent variable is measured far away from $x = 0$. A possible example of what could happen is shown in Figure 3.11.

3.8.2 Estimation and Testing for Regression through the Origin

When $\beta_0 = 0$, the least-squares estimate, $\hat{\beta}_1$ of β_1 is given by minimizing

$$S = \sum_{i=1}^n (y_i - \beta_1 x_i)^2 \quad (3.161)$$

and is easily found to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \quad (3.162)$$

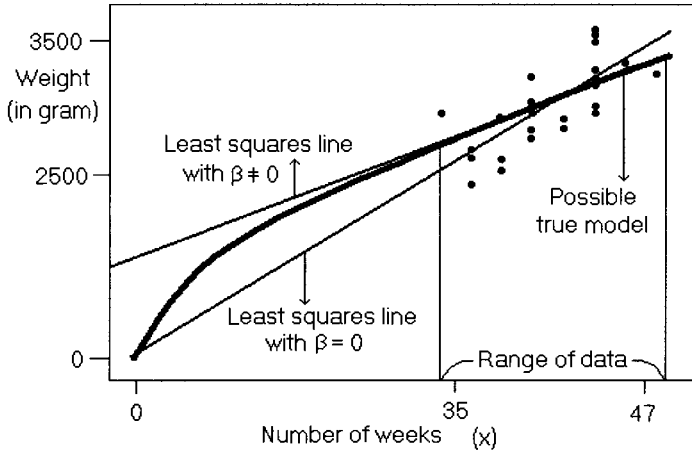


Figure 3.11: True relationship between weight and gestation period in Ex. 3.6

If the errors ε_i , $1 \leq i \leq n$, are independent $N(0, \sigma^2)$ random variables, then $\hat{\beta}_1$ is unbiased and

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}. \quad (3.163)$$

In this case $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum_{i=1}^n x_i^2)$ and

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-1} = \frac{SSE}{n-1} \quad (3.164)$$

is an unbiased estimate of σ^2 .

Additionally, SSE/σ^2 is $\chi^2(n-1)$ and independent of $\hat{\beta}_1$. The significance of the regression can be tested using the T statistic

$$T = \frac{\hat{\beta}_1}{s / \sqrt{\sum_{i=1}^n x_i^2}} = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \quad (3.165)$$

which has a t -distribution with $n-1$ degrees of freedom. A $(1-\alpha) \times 100\%$ confidence interval for β_1 is given by

$$(\hat{\beta}_1 \pm t_{n-1, \alpha/2} \hat{\sigma}(\hat{\beta}_1)) \quad (3.166)$$

From this discussion one can see that the basic theory for the case $\beta_0 = 0$ is quite similar to that for $\beta_0 \neq 0$. However, the one place where there is some difference is in the ANOVA table and in the development of a goodness of fit measure. The problem here is that the basic decomposition of the sum of squares $SST = SSR + SSE$ fails to hold in this case. The reason for this is that the sum of the residuals $\sum_{i=1}^n (y_i - \hat{y}_i)$ is not always zero in this case so that $\bar{\hat{y}} \neq \bar{y}$ which is required to prove (3.143). Thus, an R^2 value cannot be defined via (3.138) so we need to modify the ANOVA table and the ANOVA approach to testing.

We begin with a new decomposition of a sum of squares, which is valid both in the case $\beta_0 = 0$ and $\beta_0 \neq 0$. We give a proof for $\beta_0 = 0$.

Theorem 3.8 *In the linear regression model (3.160) with $\beta_0 = 0$*

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (3.167)$$

Proof. Writing $y_i = (y_i - \hat{y}_i) + \hat{y}_i$ and squaring gives

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i. \quad (3.168)$$

Now from the least square equation $\partial S / \partial \beta_1 = 0$, we get

$$\sum_{i=1}^n (y_i - \hat{\beta}_1 x_i) x_i = 0, \quad (3.169)$$

and multiplying (3.169) by $\hat{\beta}_1$ gives $(\hat{y}_i = \hat{\beta}_1 x_i)$

$$\sum_{i=1}^n (y_i - \hat{y}_i) \hat{y}_i = \sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i = 0. \quad (3.170)$$

Thus the last term in (3.168) is zero and (3.167) holds. ■

If we now use $\sum_{i=1}^n y_i^2$ as a measure of the variability of the data, then a reasonable analogue of R^2 in this case seems to be

$$R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} \quad (3.171)$$

and

$$1 - R^2 = 1 - \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n y_i^2}. \quad (3.172)$$

If we now define the F statistic by

$$F = \frac{(n-1) R^2}{1 - R^2} \quad (3.173)$$

then from (3.172)

$$F = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-1)} = \frac{\hat{\beta}_1^2 \sum_{i=1}^n x_i^2}{s^2} = T^2 \quad (3.174)$$

so that with this definition the relation between R^2 , F and T^2 is the same as when $\beta_0 \neq 0$.

However, this definition of R^2 is generally not used in practice, since it does not allow for a direct comparison of the goodness of fit of models with and without an intercept. To see this suppose we wish to compare the fit of a model with $\beta_0 \neq 0$ with one where $\beta_0 = 0$. It appears reasonable to choose as the “better” model the one with the larger R^2 .

However, when $\beta_0 = 0$

$$R^2 = 1 - \frac{SSE}{\sum_{i=1}^n y_i^2} \tag{3.175}$$

and since one generally has $\sum_{i=1}^n (y_i - \bar{y})^2 < \sum_{i=1}^n y_i^2$ the R^2 value in the zero intercept model can be considerably larger than the R^2 value when $\beta_0 \neq 0$ even if the SSE s are comparable in both cases.

Because of this problem, the definition of an appropriate R^2 value when $\beta_0 = 0$ has generated some controversy and a number of alternative definitions have been given in [47]. Perhaps a satisfactory choice is to use R^2 as the square of the correlation coefficient between $\mathbf{y} = (y_1, y_2, \dots, y_n)$ and $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ because this property holds when $\beta_0 \neq 0$. We will use this definition in all further discussions of the zero intercept model.

Another way of comparing the fit of the zero intercept model with that when $\beta_0 \neq 0$ is to use s^2 : The model with the smaller value being preferred as the one with the better fit [47].

Summarizing, the ANOVA table for the zero intercept model is given in the format shown below.

Table 3.14 ANOVA Table for the Model with $\beta_0 = 0$

Source	df	Sum of Squares	Mean Squares	F
Regression	1	$SSR = \sum_{i=1}^n \hat{y}_i^2$	$MSR = SSR/1$	$\frac{SSR}{s^2}$
Residual	$n - 1$	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n - 1} = s^2$	
Total	n	$SST = \sum_{i=1}^n y_i^2$		
$R^2 = \rho^2(\mathbf{y}, \hat{\mathbf{y}})$				

Example 3.18 In Example 3.6 (Birth weight data) we considered the relation between birth weight and gestation period for newborn babies. The fitted model was

$$\hat{y} = -1393.0 + 113.20x.$$

Further analysis gives the ANOVA table:

Table 3.15 ANOVA Table for Birth weight Data

Source	df	Sum of Squares	Mean Squares	F
Regression	1	973, 295	973, 295	25.05
Residual	22	854, 707	38, 850	
Total	23	1, 828, 003		
$R^2 = 0.532$				

From the F value we see that the regression is significant at level $< 0.01\%$. However, the t value for $\hat{\beta}_0$ is -1.60 which is not significant at the 5% level ($p = 0.125$).

In addition, $\hat{\beta}_0$ has a large negative value which gives an unphysical value if we extrapolate to $x = 0$. Although $x = 0$ is far out of the range of the observed data, it is reasonable to consider refitting the data with a line through the origin. Doing this gives $\hat{\beta}_1 = 77.130$ and $t_1 = 71.59$ which is significant at less than the 0.01% level. Moreover, $\hat{\sigma}(\hat{\beta}_1) = 1.077$ for this model compared to 22.62 for the intercept model. Hence, the slope is estimated much more precisely than when the intercept is included.

For the intercept model $R^2 = 0.532$ and $s = 197.1$ while for the zero-intercept model $s = 203.6$ and $\hat{\rho}^2(\mathbf{y}, \hat{\mathbf{y}}) = 0.533$. Thus both models appear to fit the data equally well, but the greater precision of the slope for the zero-intercept model suggests it might be a better choice for prediction.

3.8.3 Prediction

Once one has fitted the regression model to the data it is often used to estimate values of y at values of x which were not in the original data set. In fact, making such predictions is probably the most likely reason one attempts to fit a model in the first place.

There are two types of prediction we consider. First, we can predict a value of the mean response $\mu_{x_0} = E(Y_{x_0})$ at a point x_0 . Second, we can predict the value of a new observation $Y_{x_0} \equiv Y_0$ taken at x_0 . For both of these we use the point estimate

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \quad (3.176)$$

However, interval estimates are different in each case. Since $\mu_{x_0} = \beta_0 + \beta_1 x_0$ is a parameter, we can obtain confidence intervals for it, at least under our standard normality assumptions. To do this we begin by obtaining a formula for $Var(\hat{Y}_0)$.

Using the formulas for $(\hat{\beta}_0, \hat{\beta}_1)$, $\hat{Y}_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$ and the fact that $Cov(\bar{Y}, \hat{\beta}_1) = 0$, as shown in Theorem 3.1 gives

$$\begin{aligned} Var(\hat{Y}_0) &= Var(\bar{Y}) + (x_0 - \bar{x})^2 Var(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \end{aligned} \quad (3.177)$$

To estimate $Var(\hat{Y}_0)$ we replace σ^2 in (3.177) by s^2 giving

$$\hat{\sigma}^2(\hat{Y}_0) = s^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \quad (3.178)$$

The square root of $\hat{\sigma}^2(\hat{Y}_0)$, $\hat{\sigma}(\hat{Y}_0)$, is usually called the *standard error of prediction* at $x = x_0$.

Now if the errors are independent $N(0, \sigma^2)$, then,

$$\hat{Y}_0 \sim N \left(\mu_{x_0}, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \right), \quad (3.179)$$

so that

$$\frac{\hat{Y}_0 - \mu_{x_0}}{\sigma(\hat{Y}_0)} \sim N(0, 1). \quad (3.180)$$

Since \hat{Y}_0 and s are independent random variables in this case, the random variable

$$T = \frac{\hat{Y}_0 - \mu_{x_0}}{\hat{\sigma}(\hat{Y}_0)} \quad (3.181)$$

has a t distribution with $n - 2$ degrees of freedom. Thus, using standard manipulations, a $(1 - \alpha) \times 100\%$ confidence interval for μ_{x_0} is given by

$$\hat{Y}_0 \pm t_{n-2, \alpha/2} \hat{\sigma}(\hat{Y}_0). \quad (3.182)$$

One can see from (3.179) that we get the shortest confidence intervals for $x_0 = \bar{x}$ and as $|x_0 - \bar{x}|$ increases, these confidence intervals increase in width. In particular, the farther away we are from the region where the original data are taken, the less reliable are our predictions. Since the true model may not even be valid outside the region where data have been taken, one must be quite cautious in using the model to predict Y outside of the interval $[\min x_i, \max x_i]$.

Interval estimates for Y_{x_0} are not confidence intervals since Y_{x_0} is not a parameter. These estimates are referred to as *prediction intervals*. To obtain these, we first consider the variance of

$$Y_{x_0} - \hat{Y}_{x_0}.$$

If the new observation Y_{x_0} is independent of Y_i , $1 \leq i \leq n$, then

$$\begin{aligned} \text{Var}(Y_{x_0} - \hat{Y}_{x_0}) &= \text{Var}(Y_{x_0}) + \text{Var}(\hat{Y}_{x_0}) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \end{aligned} \quad (3.183)$$

To estimate $\text{Var}(Y_{x_0} - \hat{Y}_{x_0})$ we replace σ^2 with s^2 and hence $\sigma(Y_{x_0} - \hat{Y}_{x_0})$ is estimated by

$$s_p = s \sqrt{1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}}}. \quad (3.184)$$

Thus, under our standard normality assumptions

$$T = \frac{Y_{x_0} - \hat{Y}_{x_0}}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \quad (3.185)$$

will have a t distribution with $n - 2$ degrees of freedom. Hence

$$P \{ -t_{n-2, \alpha/2} \leq T \leq t_{n-2, \alpha/2} \} = 1 - \alpha \quad (3.186)$$

and from this we find that

$$\hat{Y}_{x_0} - t_{n-2, \alpha/2} s_p \leq Y_{x_0} \leq \hat{Y}_{x_0} + t_{n-2, \alpha/2} s_p \quad (3.187)$$

with probability $1 - \alpha$. The interval

$$\hat{Y}_{x_0} \pm t_{n-2, \alpha/2} s_p \quad (3.188)$$

is called a $(1 - \alpha) \times 100\%$ *prediction interval* for Y_{x_0} . The same comments regarding the reliability of prediction of a new observation hold as before.

One can see from (3.183) and (3.185) that the reliability of prediction increases as the sample size n increases and as the range of the x 's, measured by S_{xx} increases, while the reliability decreases as $|x_0 - \bar{x}|$ increases. If one can choose the x_i 's a priori, one can increase the reliability of prediction by choosing the independent variable well spread out. On the other hand, as we noted previously, this tends to inflate R^2 , leading perhaps to a poorer fit. This is a fundamental contradiction in regression analysis. A good fit may not result in good prediction, while good predictions can be made from less reliable fits.

Finally, we note that these results hold strictly only if $\varepsilon_i, 1 \leq i \leq n$, are independent $N(0, \sigma^2)$. However, because under very general conditions the errors, $(\hat{\beta}_0, \hat{\beta}_1)$ are asymptotically normal, the confidence intervals for $E(Y_{x_0})$ will be valid for large n . But the prediction intervals depend on the normality of the errors even for large values of n , so may not be valid when normality is violated.

Example 3.19 In the drug response example, Example 3.2, find a 95% confidence interval for the mean reaction time if the percentage of drug in the blood stream is 6%.

Here $x_0 = 6$, so that the estimated prediction variance is

$$\hat{\sigma}^2(\hat{Y}_6) = s^2 \left[\frac{1}{5} + \frac{(6-3)^2}{10} \right] = \frac{1.1}{3} \left[\frac{1}{5} + \frac{9}{10} \right] = 0.4033$$

and the standard error of prediction at $x = 6$ is

$$\hat{\sigma}(\hat{Y}_6) = \sqrt{0.4033} = 0.6351.$$

Hence, the estimated mean response is

$$\hat{Y}_6 = -0.1 + 0.7 \times 6 = 4.1$$

so that a 95% confidence interval for \hat{Y}_6 is given by

$$4.1 \pm t_{3, 0.025} \hat{\sigma}(Y_6) = 4.1 \pm 3.182 \times 0.6361 = (2.079, 6.121).$$

A 95% prediction interval at $x = 6$ is given by

$$4.1 \pm 3.182 \sqrt{\hat{\sigma}^2(Y_6) + 1} = 4.1 \pm 3.182 \times 0.8774 = (1.308, 6.896).$$

Example 3.20 (Clark County population data) Many economic and social activities require the accurate prediction of population sizes. For example, businesses need to be able to estimate the size of a market for a product and governments need good population estimates in order to plan for schools, roads and the allocation of funds for social programs. In the United States the importance of this was recognized by the founding fathers and the constitution requires that a census be conducted every ten years.

These problems are particularly acute in the authors' home town of Las Vegas, Nevada. Las Vegas is located in Clark County, Nevada which for many years has been the fastest growing county in the United States. (As a point of interest, the famous "*Las Vegas Strip*" does not belong to the City of Las Vegas, it belongs to Clark County.) To plan for the future it is important to be able to make accurate predictions of the population.

The simplest approach seems to be to use past census data to model the growth and use this model to predict into the future. In Table 3.16 we show the census data for the years 1920-1980 and a scatter plot is shown in Figure 3.12. That plot shows steady nonlinear growth over that period.

Table 3.16 Clark County Population

Obs. No. (i)	Year (x_i)	Population (y_i)
1	1920	4,859
2	1930	8,539
3	1940	16,414
4	1950	48,589
5	1960	127,016
6	1970	273,288
7	1980	463,087

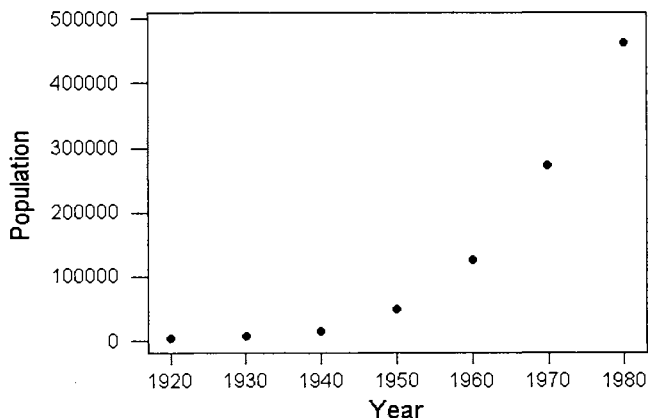


Figure 3.12: Scatter plot for Clark County population data

Although the trend is clearly nonlinear we begin our analysis by assuming a linear model

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (3.189)$$

where Y is the population and x is time coded from 0-6.

This line was fit by least squares and the results given below.

$$\begin{aligned} \hat{\beta}_0 &= -81,331, & t_0 &= -1.40, & p &= 0.22, \\ \hat{\beta}_1 &= 71,957, & t_1 &= 4.47, & p &= 0.007, \\ R^2 &= 0.80, \text{ and } & F &= 19.96. \end{aligned}$$

From these results we see that the slope β_1 is significantly different from zero and the model explains 80% of the variability in the data. However, there are some obvious problems with the estimated populations, which are shown in Table 3.17.

Table 3.17 Data, Fitted Values for Example 3.20

Year (i)	Population (y_i)	Fitted value (\hat{y}_i)	Residual ($\hat{\varepsilon}_i = y_i - \hat{y}_i$)
1920	4,859	-81,331	86,190
1930	8,539	-9,397	17,936
1940	16,414	62,584	-46,170
1950	48,589	134,541	-85,952
1960	127,016	206,498	-79,482
1970	273,288	278,455	-5,167
1980	463,087	350,412	112,675

Notice that the estimated populations in 1920 and 1930 are negative - obviously a non physical result - but the model appears to give more reasonable results at later times.

As a consequence, we consider using the model to predict the population in 1990 ($x = 7$). Using (3.176) we get

$$\hat{y}_{(1990)} = -81,331 + 71957(7) = 422,368$$

and 95% confidence and prediction intervals are given by

- (a) 95% confidence interval: (237,184, 607,554),
- (b) 95% prediction interval: (135,482, 709,256).

Even though the fit is significant, it does not predict very well, since the 1990 census gave the population of Clark County as 768,203 which lies outside the 95% prediction interval.

To improve the predictive ability of the model we consider finding one that fits the population data better. From the scatter plot it appears that the population grows exponentially, so we consider a model of the form

$$y = \gamma_0 \exp(\beta_1 x) \quad (3.190)$$

where β_1 is the *growth rate*. Since (3.190) is not linear in (γ_0, β_1) , the theory developed so far is not immediately applicable. However, taking logarithms in (3.190) gives

$$\log y = \log \gamma_0 + \beta_1 \log x \quad (3.191)$$

If we let $y' = \log y$, $\beta_0 = \log \gamma_0$ and $x' = \log x$, then

$$y' = \beta_0 + \beta_1 x'. \quad (3.192)$$

So (3.192) represents a linear relation between (x', y') . The scatter plot of (x', y') is shown in Figure 3.13 and the points fall almost on a straight line. To test this, we fitted a least squares line to (x', y') and the results are given below.

$$\begin{aligned} \hat{\beta}_0 &= 8.3379, & t_0 &= 67.07, & p &< 10^{-3}, \\ \hat{\beta}_1 &= 0.80896, & t_1 &= 23.46, & p &< 10^{-3}. \end{aligned}$$

Table 3.18 ANOVA Table for Clark County Population Data

Source	df	Sum of Squares	Mean Squares	F
Regression	1	18.324	18.324	550.46
Residual	5	0.166	0.033	
Total	6	18.490		
$R^2 = 0.991$				

From this we see that the linearized model gives an almost perfect fit to the data and is significant at $< 0.01\%$ level. One would expect that such a good fitting model would be a useful predictor. To check this we again use the model to predict the 1990 population. Taking $x' = \log 7$ in (3.192) the estimated values of y' and confidence and prediction intervals are:

$$\hat{y}'_{(1990)} = 14.004, \quad \hat{\sigma} \left(\hat{y}'_{(1990)} \right) = 0.1545$$

- (a) 95% confidence interval: (13.6041, 14.3971),
- (b) 95% prediction interval: (13.3863, 14.6148).

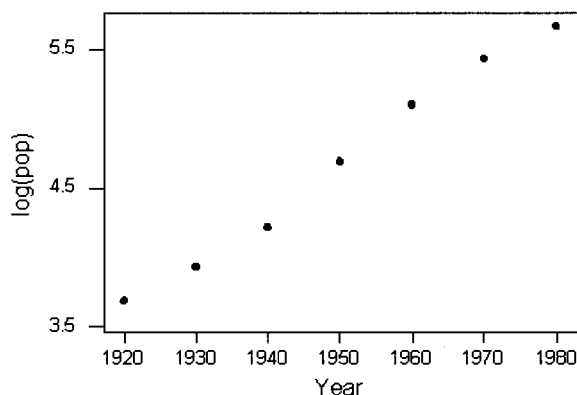
To compare these predictions with the linear model we exponentiate \hat{y}' and the end points of the intervals to get.

$$\hat{y}_{(1990)} = 1,209,842$$

- (a) 95% confidence interval: (809,360, 1,788,880),
- (b) 95% prediction interval: (651,023, 2,223,960).

Now note that compared to the true population 768,203 the predicted population is in greater error than given by the worse fitting linear model, however it does lie in a 95% prediction interval. It appears then, that the better fitting model is not necessarily the better predictor. Because we are predicting ten years beyond the range of the data we have little reason to believe that the model (3.189) and (3.190) is valid there.

As a last comment, these data illustrate that one should take with a “grain of salt” those predictions one is always seeing in the popular media. Prediction errors can be very large, but they are rarely published, e.g., the recent budget surplus predictions.

Figure 3.13: Scatter plot for Clark County population, $\log Y$

3.9 Assessing Model Validity

So far we have developed the theory of the simple linear regression model under the assumption that we know that the true model for $E(Y_x)$ is linear. Although this knowledge may be available a priori, perhaps from physical and/or theoretical considerations, in most practical situations we will not know whether the simple linear regression model is true, and an important part of the analysis is to see whether such a model is tenable. Since the violation of linearity will generally (but not always) invalidate our previous analysis concerning (β_0, β_1) it is desirable to have tests that would check for linearity before any further analysis is done. For the most part we will only have the data $\{(x_i, y_i)\}_{i=1}^n$ to work with, so they will have to form the basis for any analysis.

There are a number of procedures that can be used in this regard:

- (i) Examine the scatter plot of $\{(x_i, y_i)\}$. Pronounced curvature in the plot such as that shown in Figure 3.12 suggests that a quadratic model

$$Y_x = \beta_0 + \beta_1 x_i + \beta_2 x^2 + \varepsilon_i \quad (3.193)$$

might be a more appropriate choice than a linear model. However, the scatter plot may be confusing, or even misleading (see Example 5.13) if the departures from the simple linear model results from missing variables, rather than curvature in x . For example, Figure 3.14 suggests that there is a slight upward trend as shown by the least squares line superimposed on the scatter plot. But it is not clear if the large fluctuations are just random error, or due to some other factors which have not been accounted for. In this case it can be shown (see Example 3.21) that the fluctuations can be much better explained in terms of seasonal factors, rather than random errors. Of course, once one is aware of the source of the data, such a hypothesis is reasonable.

- (ii) Examination of the test statistics associated with a preliminary least squares fit to the data. For example, a small value of R^2 along with an apparently significant t

value for $\hat{\beta}_1$ generally suggests that the true model contains variables other than x . On the other hand, a large value of R^2 and significant t values does not of itself imply that the model is linear.

- (iii) Residual plots from the least squares fit are another effective diagnostic tool for examining the validity of the linearity assumption. Since the residuals estimate what variability remains in the data after the linear part in x has been removed, it is reasonable to expect that their values would be useful in detecting departures from linearity. There are a variety of plots that are commonly recommended for doing this, and these will be explored later in this chapter and in more detail in Chapter 6.

Example 3.21 (Jewelry sales data [123]) Suppose we wish to examine how department store sales of jewelry increase over time. Table 3.19 shows quarterly sales from 1957 to 1960. When we plot quarterly sales against time as in Figure 3.14, we note that sales shoot up every fourth quarter because of the Christmas and holiday seasons.

Table 3.19 Jewelry Sales Data		
Year	Quarter	Sales (in \$100,000)
1957	1	36
	2	44
	3	45
	4	106
1958	1	38
	2	46
	3	47
	4	112
1959	1	42
	2	49
	3	48
	4	118
1960	1	42
	2	50
	3	51
	4	118

After we fit the model $Y_t = \beta_0 + \beta_1 t + \varepsilon$, ($t = \text{time}$) we obtained the results as follows:

$$\begin{aligned} \hat{\beta}_0 &= 45.68, & t_0 &= 2.83, & p &= 0.013, \\ \hat{\beta}_1 &= 1.921, & t_1 &= 1.15, & p &= 0.269, \\ R^2 &= 0.087, \text{ and } & F &= 1.33. \end{aligned}$$

From these results we see that the hypothesis $H_0 : \beta_1 = 0$ is not rejected ($p > 0.05$) and the fitted line only explains 8.7% of the variability of sales in the data. Also since the F value is 1.33 ($< F_{1,14,0.05} = 4.60$) the estimated regression model, which is superimposed in Figure 3.14, seems to be poor. Nevertheless, it needs to be noted that the estimated slope, $\hat{\beta}_1$ is biased due to seasonality factors, which leads us to introduce some other regressor variable(s) in Chapter 7 to take seasonal factors into account.

Since scatter and residual plots can be quite subjective, it would be helpful to have a more objective analytical approach for assessing the linearity of the model. Unfortunately

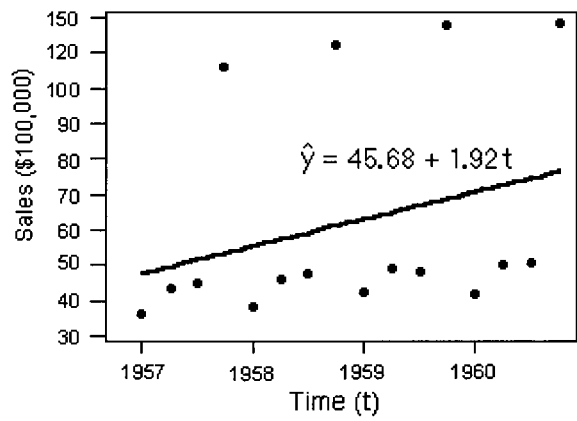


Figure 3.14: Scatter plot and fitted line of jewelry sales data

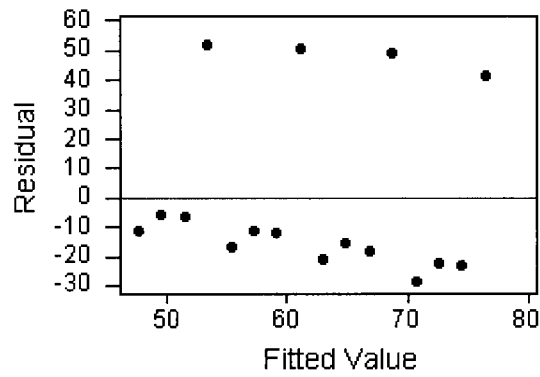


Figure 3.15: Scatter plot of residuals versus fitted values

few tools seem to be available for doing this, and for most data, particularly those from unplanned experiments, the techniques in (i)-(iii), along with judgement, are those most often used in practice. On the other hand for designed experiments, of the type that occur in industrial situations or clinical situations, there is an analytic test, the “*lack of fit*” test that is widely advocated.

3.9.1 The Lack of Fit Test (LOFT)

The basic idea of the LOFT is to obtain an independent estimate of σ^2 , other than that given by s^2 . These two estimates can then be compared in a suitable test to determine whether the linearity assumption is viable.

To be more formal, suppose we wish to decide which one of the two models

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad 1 \leq i \leq n, \quad (3.194)$$

or

$$Y_i = \beta_0 + \beta_1 x_i + \eta_i + \varepsilon_i, \quad 1 \leq i \leq n, \quad (3.195)$$

best explains our data. Here, η_i represents the departure of the model from the simple linear regression model. As before, we assume in both cases that $\{\varepsilon_i\}_{i=1}^n$ are independent $N(0, \sigma^2)$. Thus, we wish to test

$$H_0 : \eta_i = 0, \quad 1 \leq i \leq n, \quad (3.196)$$

against

$$H_1 : \text{at least one } \eta_i \neq 0, \quad 1 \leq i \leq n. \quad (3.197)$$

To construct a test of H_0 consider the estimate of σ^2 given by s^2 when the data have been fit by least squares. If H_0 is true then as shown in Theorem 3.2 $E(s^2) = \sigma^2$, while if H_1 is true it can be shown that

$$E(s^2) = \sigma^2 + \frac{1}{n-2} \sum_{i=1}^n \left[E(Y_i) - E(\hat{Y}_i) \right]^2 \quad (3.198)$$

where \hat{Y}_i is the least squares estimator of Y_i . The term $E(Y_i) - E(\hat{Y}_i)$ represents the *bias* in estimating Y_i by least squares if (3.195) is the true model and can be shown to be a function of η_i . Of course if H_0 is true, then $E(Y_i) = E(\hat{Y}_i)$ and $E(s^2) = \sigma^2$ as we know. From (3.198) we see that if H_1 is true then s^2 will tend to be larger than σ^2 . If σ^2 were known, then we could reject H_0 if the ratio s^2/σ^2 was sufficiently large. In general, however, σ^2 will not be known, so for this procedure to be useful we must have another, independent estimate of σ^2 . In situations where we have more than one observation at some of the points x_i then such an estimate may be obtained as follows.

Suppose that we have m distinct x_j -values, $1 \leq j \leq m$ and at each point x_j we have n_j observations y_{ij} , $1 \leq i \leq n_j$ where at least one $n_j > 1$. Now for each x_j we can obtain an estimate of σ^2 by the standard formula

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \quad (3.199)$$

where

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}. \quad (3.200)$$

At points where $n_j = 1$ we take $\hat{\sigma}_j^2 = 0$. Using (3.200) and pooling the estimates $\hat{\sigma}_j^2$ of σ^2 in the usual way we get a further estimate of σ^2

$$\hat{\sigma}_p^2 = \frac{1}{n-m} \sum_{j=1}^m (n_j - 1) \hat{\sigma}_j^2 \quad (3.201)$$

where $n = \sum_{j=1}^m n_j$ is the total number of observations. If the errors are independent $N(0, \sigma^2)$, then $\hat{\sigma}_p^2 / \sigma^2$ is $\chi^2(n-m)$.

The quantity $(n-m) \hat{\sigma}_p^2$ is called the *pure error sum of squares*, whence the subscript p on $\hat{\sigma}_p^2$.

Using $\hat{\sigma}_p^2$ we can now reject H_0 if

$$Q = s^2 / \hat{\sigma}_p^2 \quad (3.202)$$

is sufficiently large. To complete the test, we will need to obtain the distribution of (3.202) and for this a slight modification of (3.202) is necessary.

If we subtract $(n-m) \hat{\sigma}_p^2$ from $(n-2) s^2$ we get a new sum of squares

$$SS_{LOF} = (n-2) s^2 - (n-m) \hat{\sigma}_p^2 \quad (3.203)$$

called the *lack of fit sum of squares* which is an estimate of the bias term in (3.202).

Using this in (3.202) gives

$$Q = \frac{[SS_{LOF} + (n-m) \hat{\sigma}_p^2] / (n-2)}{\hat{\sigma}_p^2}. \quad (3.204)$$

Rearranging (3.204) we see that we can reject H_0 if

$$SS_{LOF} / \hat{\sigma}_p^2 \quad (3.205)$$

is sufficiently large.

If the errors are $N(0, \sigma^2)$ then SS_{LOF} / σ^2 has a χ^2 -distribution with $(n-2) - (n-m) = m-2$ degrees of freedom. Therefore an equivalent test can be based on the F -ratio

$$F = \frac{[SS_{LOF} / (m-2)]}{\hat{\sigma}_p^2} \quad (3.206)$$

and we reject H_0 at level α if

$$F > f_{m-2, n-m, \alpha} \quad (3.207)$$

where $P\{F > f_{m-2, n-m, \alpha}\} = \alpha$.

If (3.207) is satisfied we say that the model suffers from *lack of fit*, so that the simple linear model appears to be untenable. In this case the least squares analysis discussed in Section 3.2 should not be used, and further analysis to determine the source(s) of the bias in the model should be performed. Often, the cause of the problem is that other

explanatory variables have been omitted, and if the variables which have been left out can be ascertained, then further fitting and analysis via multiple regression techniques can often deal with the problem. This will be taken up in Chapter 5.

If H_0 is accepted, then the linearity assumption is tenable and further analysis of the model such as significance testing for $\beta_1 = 0$, is legitimate, and one may proceed as in Sections 3.3-3.7. This assumes, of course, that all the other assumptions made there are valid. These should be tested as well and is usually done through residual plots.

Before giving a numerical example we summarize the procedure for using the lack of fit test in fitting a simple linear regression model with independent $N(0, \sigma^2)$ errors.

- (i) Fit the model $y = \beta_0 + \beta_1x$ to the data using least squares and construct the usual ANOVA table, but do not use the F -ratio to test for the significance of the regression.
- (ii) Calculate the pure error sum of squares (SS_{PE}) and subtract it from SSE to get the lack of fit sum of squares (SS_{LOF}).
- (iii) Form the F -ratio

$$F = \frac{SS_{LOF}/(m-2)}{\hat{\sigma}_p^2}$$

(3.208)

and compare this to the tabulated $f_{m-2,n-m,\alpha}$ value (typically $\alpha = 0.05$).

- (iv) If $F > f_{m-2,n-m,\alpha}$ then the model displays lack of fit and the simple linear model is not tenable. In this case, further analysis will be necessary to determine an appropriate model.
- (v) If there is no significant lack of fit, then one can entertain the model (3.1) as plausible and proceed to test for the significance of $\hat{\beta}_1$, etc. This procedure is conveniently carried out by expanding the ANOVA table as shown below.

Table 3.20 ANOVA Table with Lack of Fit Analysis

Source	df	Sum of Squares	Mean Squares	F
Regression	1	SSR	$MSR = SSR$	$\frac{MSR}{MSE}$
Residual	$n - 2$	SSE	$MSE = \frac{SSE}{n-2}$	
Lack of Fit	$m - 2$	SS_{LOF}	$MS_{LOF} = \frac{SS_{LOF}}{m-2}$	$\frac{MS_{LOF}}{MS_{PE}}$
Pure Error	$n - m$	SS_{PE}	$MS_{PE} = \frac{SS_{PE}}{n-m}$	
Total	$n - 1$	SST		
$R^2 = \frac{SSR}{SST}$				

Note: m = number of distinct x values

Example 3.22 In a manufacturing process, the effect of temperature on the color of a finished product was determined experimentally. The data collected were as follows:

Table 3.21 Color Data			
Temperature (x)	Color (y)	Temperature (x)	Color (y)
460	0.3	430	0.6
450	0.3	420	0.6
440	0.4	410	0.7
430	0.4	400	0.6
420	0.6	420	0.6
410	0.5	410	0.6
450	0.5	400	0.6
440	0.6		

Determine if a linear model is plausible for this data.
We begin by making a scatter plot of (x, y) . This is shown below in Figure 3.16.

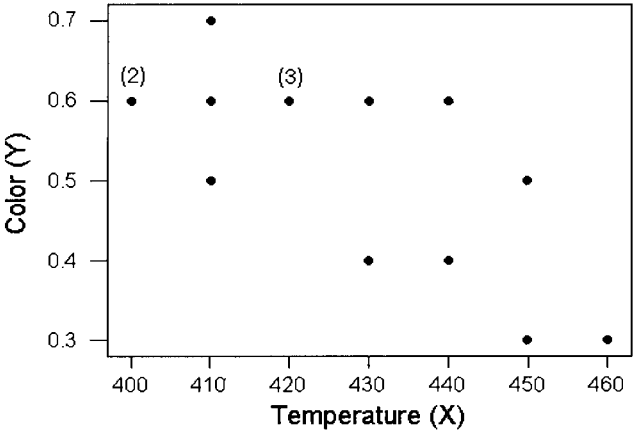


Figure 3.16: Scatter plot for color data

From the scatter plot there appears to be an overall downward trend in the data. Since we have repeat points we can use the lack of fit test to test analytically if the linear hypothesis is tenable.

We begin by fitting the least squares line to the data. This gives the fitted line

$$y = 2.536247 - 0.0047177x.$$

Using this, we find that $SSE = 0.098938$. To calculate the F -ratio for lack of fit we need to compute the pure error sum of squares. These calculations are summarized below.

Table 3.22 Calculations for SS_{PE}

Temperature (x)	SSE_j	df
400	0	1
410	0.02	2
420	0	2
430	0.02	1
440	0.02	1
450	0.02	1
Total	0.08	8

This information is summarized in the following ANOVA table.

Table 3.23 ANOVA Table with Lack of Fit Analysis

Source	df	Sum of Squares	Mean Squares	F
Regression	1	0.110395	0.110395	14.51
Residual	13	0.098938	0.007611	
Lack of Fit	8	0.018938	0.010000	0.38
Pure Error	5	0.080000	0.003788	
Total	14	0.209333		
$R^2 = 0.5273$				

From Table A.4 we find that $f_{5,8,0.05} = 3.69$ and the observed lack of fit $F = 0.38 < 3.69$ so that the hypothesis of lack of fit is rejected and we accept the linear model as tenable.

In this case, assuming that the errors are independent $N(0, \sigma^2)$, the F test for the significance of the regression is appropriate. Again from Table A.4 we find that $f_{1,3,0.05} = 4.67$ and the observed F ratio 14.51 exceeds this. Thus, we can conclude that $\beta_1 \neq 0$ at the 5% level of significance.

Keep in mind that even if lack of fit is rejected and the regression is found to be significant, this does not mean that (3.194) is the true model. The best that we can say is that the data appear to be consistent with this hypothesis. Further investigation might prove otherwise.

When the lack of fit test is not appropriate, then other means are needed to check model adequacy. Of course we can use T and F statistics as before, but they may be significant even if (3.1) is not the true model. Our previous analysis suggests that residuals $\hat{\varepsilon}_i$ should be useful in this respect. Intuitively, if (3.1) is the true model, then we would expect that the residuals should behave like a random sample of size n of a $N(0, \sigma^2)$ random variable. To the extent that the residuals do not appear to behave that way, this will provide evidence of the inadequacy of (3.1). The most common way of doing this is to make various graphical plots of the residuals and examine these plots for apparent deviations from the model assumptions. In this section we will examine a number of basic plots and in Chapter 6 these ideas will be extended to multiple regression models. Before doing this, we state some additional properties of $\hat{\varepsilon}_i$, $1 \leq i \leq n$.

Theorem 3.9 *Let $\hat{\varepsilon}_i$, $1 \leq i \leq n$, be the residuals from the least squares fit of (3.1). Then,*

$$(i) \ E(\hat{\varepsilon}_i) = 0, 1 \leq i \leq n;$$

$$(ii) \text{Var}(\hat{\varepsilon}_i) \equiv \sigma_{\hat{\varepsilon}_i}^2 = \sigma^2 \left\{ 1 - \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \right\} \simeq \sigma^2 \text{ for large } n;$$

$$(iii) \text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_i)(\bar{x} - x_j)}{S_{xx}} \right];$$

(iv) If the errors, $\varepsilon_i, 1 \leq i \leq n$, are $N(0, \sigma^2)$, then

$$\hat{Z}_i = \frac{\hat{\varepsilon}_i}{\sigma_{\hat{\varepsilon}_i}} \sim N(0, 1); \quad (3.209)$$

$$(v) \text{Cov}(\hat{\varepsilon}_i, \hat{Y}_i) = 0, 1 \leq i \leq n.$$

Proof. (i) By definition, $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ so that $E(\hat{\varepsilon}_i) = E(Y_i - \hat{Y}_i) = E(Y_i) - E(\hat{Y}_i)$. But, $E(\hat{Y}_i) = E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = E(\hat{\beta}_0) + x_i E(\hat{\beta}_1) = \beta_0 + x_i \beta_1 = E(Y_i)$ since $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased. Thus $E(\hat{\varepsilon}_i) = 0$, as required.

(ii) $\text{Var}(\hat{\varepsilon}_i) = \text{Var}(Y_i - \hat{Y}_i) = \text{Var}(Y_i) + \text{Var}(\hat{Y}_i) - 2\text{Cov}(Y_i, \hat{Y}_i)$. From (3.69) and (3.74) $\text{Var}(\hat{Y}_i) = \sigma^2 [1/n + (x_i - \bar{x})^2 / S_{xx}]$ and $\text{Cov}(Y_i, \hat{Y}_i) = \sigma^2 [1/n + (x_i - \bar{x})^2 / S_{xx}]$. Thus,

$$\begin{aligned} \text{Var}(\hat{\varepsilon}_i) &= \sigma^2 - \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \\ &= \sigma^2 \left\{ 1 - \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \right\}. \end{aligned} \quad (3.210)$$

(iii) Now

$$\begin{aligned} \text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) &= \text{Cov}(Y_i - \hat{Y}_i, Y_j - \hat{Y}_j) \\ &= \text{Cov}(Y_i, Y_j) - \text{Cov}(Y_i, \hat{Y}_j) - \text{Cov}(Y_j, \hat{Y}_i) + \text{Cov}(\hat{Y}_i, \hat{Y}_j) \\ &= -\text{Cov}(Y_i, \hat{Y}_j) - \text{Cov}(Y_j, \hat{Y}_i) + \text{Cov}(\hat{Y}_i, \hat{Y}_j) \end{aligned} \quad (3.211)$$

since by independence $\text{Cov}(Y_i, Y_j) = 0$.

But,

$$\begin{aligned} \text{Cov}(\hat{Y}_i, \hat{Y}_j) &= \text{Cov}(\hat{\beta}_0 + x_i \hat{\beta}_1, \hat{\beta}_0 + x_j \hat{\beta}_1) \\ &= \text{Var}(\hat{\beta}_0) + (x_i + x_j) \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \text{Var}(\hat{\beta}_1). \end{aligned} \quad (3.212)$$

From Theorem 3.1 $\text{Var}(\hat{\beta}_0) = \sigma^2 [1/n + \bar{x}^2 / S_{xx}]$ and $\text{Var}(\hat{\beta}_1) = \sigma^2 / S_{xx}$. Also, (3.68) gives $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \bar{x} / S_{xx}$. Thus,

$$\begin{aligned} \text{Cov}(\hat{Y}_i, \hat{Y}_j) &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} - \frac{(x_i + x_j)\bar{x}}{S_{xx}} + \frac{x_i x_j}{S_{xx}} \right] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_i)(\bar{x} - x_j)}{S_{xx}} \right]. \end{aligned} \quad (3.213)$$

A similar calculation shows that

$$\text{Cov}(Y_i, \hat{Y}_j) + \text{Cov}(Y_j, \hat{Y}_i) = 2\sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_i)(\bar{x} - x_j)}{S_{xx}} \right] \quad (3.214)$$

so that

$$\text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_i)(\bar{x} - x_j)}{S_{xx}} \right]. \quad (3.215)$$

(iv) If ε_i is $N(0, \sigma^2)$, then $\hat{\varepsilon}_i$ is normal as well, since $\hat{\varepsilon}_i$ is a sum of the Y_i 's (show this). From (i)-(ii) $E(\hat{\varepsilon}_i) = 0$ and $\text{Var}(\hat{\varepsilon}_i) = \sigma_{\hat{\varepsilon}_i}^2$. Hence, the normalized random variables $\hat{Z}_i = \hat{\varepsilon}_i / \sigma_{\hat{\varepsilon}_i}$ are $N(0, 1)$.

(v) $\text{Cov}(\hat{\varepsilon}_i, \hat{Y}_i) = \text{Cov}(Y_i - \hat{Y}_i, \hat{Y}_i) = \text{Cov}(Y_i, \hat{Y}_i) - \text{Var}(\hat{Y}_i)$. From (3.213) and (3.214) $\text{Cov}(Y_i, \hat{Y}_i) = \text{Var}(\hat{Y}_i) = \sigma^2 \left[1/n + (x_i - \bar{x})^2 / S_{xx} \right]$ so that $\text{Cov}(\hat{\varepsilon}_i, \hat{Y}_i) = 0$ as required. ■

From (3.215) one can show that $\text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) \simeq 0$ for large n . Hence, as stated above, if the errors are independent $N(0, \sigma^2)$, then for large n the *standardized residuals* $\hat{\varepsilon}_i / \sigma_{\hat{\varepsilon}_i}$ should behave as a random sample from a $N(0, 1)$ random variable.

Of course, for practical purposes one needs an estimate of σ^2 in order to calculate \hat{Z}_i . There are a number of ways of doing this. The most well-known procedure is to estimate σ^2 by s^2 and \hat{Z}_i is estimated by

$$\hat{r}_i = \frac{\hat{\varepsilon}_i}{s \sqrt{1 - \left[1/n + (x_i - \bar{x})^2 / S_{xx} \right]}} \quad (3.216)$$

which are usually referred to as *studentized residuals* (sometimes called *internally studentized residuals*). Again, for large n we expect that \hat{r}_i should behave like a random sample from a $N(0, 1)$ random variable.

Although the ordinary residuals, $\hat{\varepsilon}_i$, $1 \leq i \leq n$, are most commonly used for plotting, there is another class of residuals which are of increasing importance in regression analysis, the *PRESS residuals*. These residuals are obtained by fitting the model (3.1) omitting the i -th data point. The resulting estimates of β_0 and β_1 are denoted by $\beta_{0(-i)}$ and $\beta_{1(-i)}$ and $E(Y_i)$ is estimated by

$$\hat{Y}_{(-i)} = \beta_{0(-i)} + x_i \beta_{1(-i)}. \quad (3.217)$$

Then, the i -th *PRESS residual* is defined by

$$\hat{\varepsilon}_{(-i)} = \hat{Y}_i - \hat{Y}_{(-i)}. \quad (3.218)$$

These can be normalized by an appropriate estimate of $\sigma_{(-i)}^2 \equiv \text{Var}(\hat{\varepsilon}_{(-i)})$ giving the *externally studentized residuals*. In contrast to the $\hat{\varepsilon}_i$, these residuals are more amenable to mathematical analysis and provide a non-graphical means for residual examination. Further discussion of these will be given in Chapter 6. For now we concentrate on using $\hat{\varepsilon}_i$ and \hat{r}_i .

3.9.2 Residual Plots

We now present a number of diagnostic residual plots which enable us to determine violations/departures in the assumptions in the simple linear regression models. These include histograms of the residuals, normal residual plots and plots of the residuals versus the fitted values \hat{y}_i , $1 \leq i \leq n$ and the independent variables x_i , $i = 1, 2, \dots, n$.

Histograms

The histogram is one of the simplest graphical summaries used to visualize the patterns of the residuals, $\hat{\varepsilon}_i$, $i = 1, 2, \dots, n$. The values of $\hat{\varepsilon}_i$ may be arranged into a frequency distribution (with appropriate size of classes of residuals) and then plotted in a histogram. The horizontal axis (center is zero) represents the range of values of the residuals and the vertical axis indicates the frequency of each class of residuals. If the shape of the depicted histogram is close to an approximately bell-shaped/mound-shaped curve, there would be no reason to suspect that the normality assumption has been violated.

Normal Probability Plots

To obtain normal probability plots we define the cumulative distribution function (cdf) of a $N(0, 1)$ random variable by

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad (3.219)$$

and define percentiles by

$$\Phi(q) = p \quad (3.220)$$

where p is a given probability. Then q , the $p \times 100$ th percentile is given

$$q = \Phi^{-1}(p) \quad (3.221)$$

where Φ^{-1} is the inverse cdf of Φ .

To make normal residual plots we order the residuals $\hat{\varepsilon}_i$, $1 \leq i \leq n$, by

$$\hat{\varepsilon}_{(1)} \leq \hat{\varepsilon}_{(2)} \leq \dots \leq \hat{\varepsilon}_{(n)} \quad (3.222)$$

where $\hat{\varepsilon}_{(i)}$ denote the ordered values of the residuals. Similarly we can order the standardized residuals \hat{r}_i , $1 \leq i \leq n$, as

$$\hat{r}_{(1)} \leq \hat{r}_{(2)} \leq \dots \leq \hat{r}_{(n)}. \quad (3.223)$$

The ordered residuals are then plotted against $\Phi^{-1}[(i - 1/2)/n]$, $1 \leq i \leq n$.

From the fact that $E[\varepsilon_{(i)}] \simeq \Phi^{-1}[(i - 1/2)/n]$ if the errors are normally distributed, then the points should lie on an approximately straight line $y = x$ (y is the axis for $\hat{\varepsilon}_{(i)}$ or $\hat{r}_{(i)}$). Such plots are also useful for detecting outliers or dubious observations. If the plot is S-shaped, it indicates that the distribution has relatively light (or short) tails. On the other hand, a heavy (or long) tailed sampling distribution tends to look like a backward S. Positively skewed distributions tend to have a J shape while it has an inverted J shape for negatively skewed distributions.

Hence, normal probability plots can be used to detect departures from the normality assumptions in (3.1).

Variable Plots

- (i) Since, asymptotically the *studentized residuals* \hat{r}_i do not depend on σ a plot of \hat{r}_i versus x_i can be used to detect deviations from the linearity assumption and/or violations of constant variance.
- (ii) From (v) of Theorem 3.9 we see that $\hat{\varepsilon}_i$ and \hat{Y}_i are uncorrelated, hence if the model (3.1) is true, a plot of $\hat{\varepsilon}_i$ or \hat{r}_i against \hat{y}_i should exhibit random scatter about a horizontal band as indicated in Figure 3.17(a). If we use \hat{r}_i , then these form an approximate random sample from a $N(0, 1)$ random so it is unlikely that $|\hat{r}_i| > 2$. Hence, most of the residuals should fall between ± 2 . Residuals that fall outside this band, may indicate discrepant observations - called *outliers* - that require further investigation (in Figure 3.17(b)). Similar statements apply to residual plots against x_i . Violations in the assumption of constant variance can also be detected using externally studentized residuals similar to the plots of \hat{r}_i against x_i or \hat{y}_i . In this case the residuals will display some systematic behavior in either \hat{y} or x such as shown in Figure 3.17(c) or Figure 3.17(d). Such patterns often occur because the error distribution is not normal, and have variances which depend on the mean. For example, a Poisson random variable X with parameter λ , $E(X) = \lambda$ and $\sigma(X) = \sqrt{\lambda} = \sqrt{E(X)}$. This suggests that $\hat{\varepsilon}_i^2 \simeq \hat{y}_i$ giving rise to the funnel pattern shown in Figure 3.17(c). Since count data often follow a Poisson distribution, such non-constant variance should be suspected for such data.

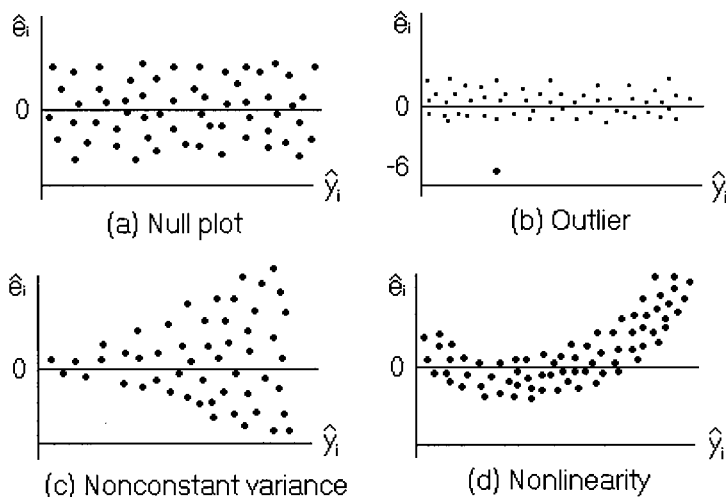


Figure 3.17: Typical residual plots

- (iii) In some instances residual plots can be used to detect violations of the independence assumption. In particular, if the independent variable is time, such as in Example 3.5 (Tractor data) or economic models (stock price, GNP, unemployment rates, etc.) then systematic patterns in the residuals may indicate *serial correlation* between the observations as indicated in Figure 3.18.

- (iv) I Charts: In these residuals are plotted in the order the observations were taken. These can be useful in identifying serial correlation in the data.

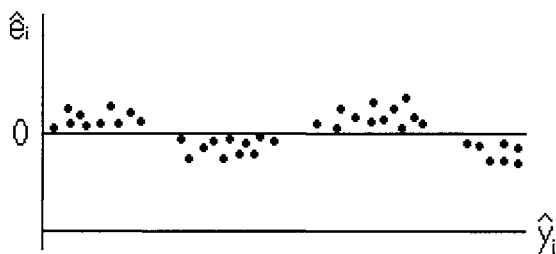


Figure 3.18: Residual plot if serial correlation exists

Example 3.23 (Tractor data) Using the Tractor data in Example 3.5, we illustrate residual plots to detect any violations in the model assumptions. In Figure 3.19, although the normal plot is roughly a straight line, the histogram indicates that the errors are not normally distributed, it looks rather uniform. The I Chart has a somewhat systematic pattern in the plot.

Example 3.24 (Drink delivery data) Using data in Table 3.5, in Figure 3.21 we show variable residual plots: histograms, normal plots of \hat{r}_i , and plots of \hat{r}_i versus \hat{y}_i . Also the plot of \hat{r}_i versus x_i (number of cases) which looks like the funnel type in Figure 3.21.

Example 3.25 (Birth weight data) We use Example 3.16 to illustrate variable plots of the residuals. Figure 3.22 shows a normal plot of \hat{r}_i , an I Chart of the residuals, a histogram of \hat{r}_i , and a plot of \hat{r}_i versus the fitted values \hat{y}_i . The normal probability plot looks reasonable. Also, Figure 3.23 shows the scatter plot of the standardized residuals \hat{r}_i versus the variable x (age).

Example 3.26 (Clark county population data) Figure 3.24 shows variable plots for the linear fit. There are normal plot of \hat{r}_i , an I Chart of residuals, a histogram of \hat{r}_i , and a plot of \hat{r}_i versus fitted values \hat{y}_i . Clearly, the normal plot does not seem to be a straight line. This indicates that the errors are not normally distributed. This can be verified in the histogram and by the plot of residuals versus fitted values.

3.10 Transformations

In developing the theory for the simple linear model we have made a number of assumptions concerning the relationship of Y to x , the errors ε_i and the accuracy of measurement of the independent and dependent variables. If some of these assumptions are not true, then the analysis outlined in Sections 3.2-3.8 may lead one to erroneous conclusions. If these violations are detected, say through residual plots or by prior information, the question arises as to what, if anything can be done to model our data.

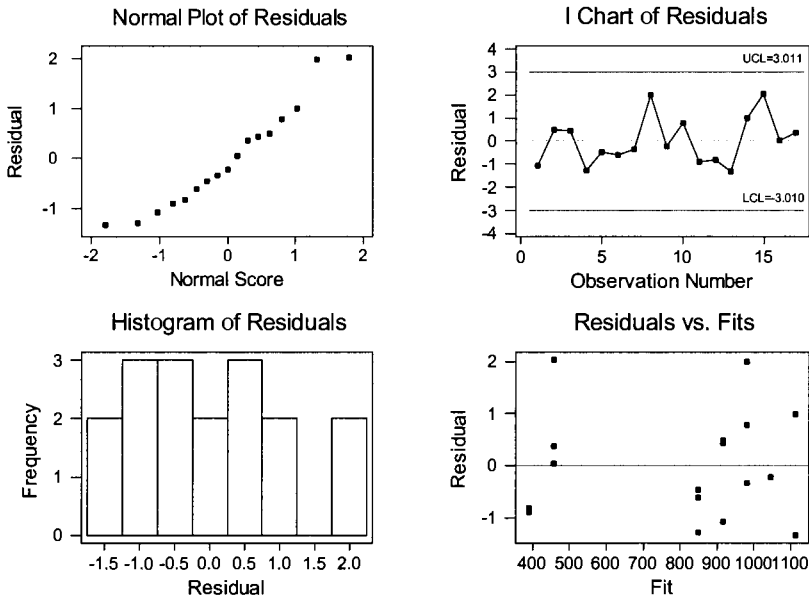


Figure 3.19: Variable plots of residuals \hat{r}_i for Tractor data

A general strategy for dealing with deviations from linearity, normality of errors and/or homoscedasticity, is to transform the variables in some fashion so that the transformed model satisfies, at least approximately, the basic assumption used in Sections 3.2-3.8. We shall pursue this approach here, and then it will be generalized to deal with multiple regression models in Chapter 6. Needless to say, trying to correct any or all of the possible violations of the basic assumptions is only part science, and one may need to rely to a great extent on judgement and knowledge of the process generating the data along with the formal tools of statistical analysis.

One should keep in mind that most of the time there will not be a one-to-one correspondence between observed model inadequacies and possible solutions, so that the remedy finally chosen may depend on other than just mathematical considerations. In some situations there may be no adequate solutions at all.

The three problems that we shall examine are:

- (i) Possible non-linearity of the mean response $E(Y_x)$ in both x and (β_0, β_1) ;
- (ii) Possible non-normality of the errors;
- (iii) Heteroscedasticity of the errors.

We focus on the cases (i)-(ii). Case (iii) will be discussed in Chapter 6.

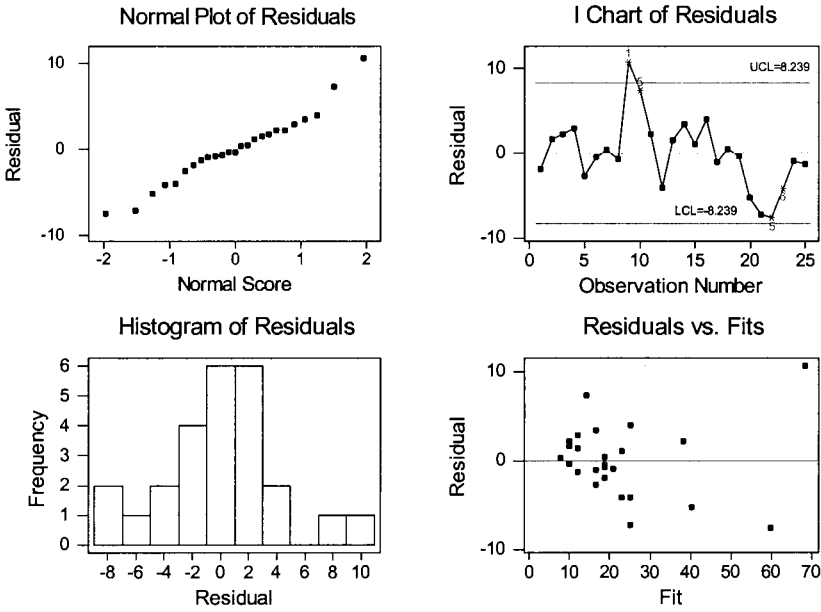


Figure 3.20: Variable plots of residuals \hat{r}_i for drink delivery data

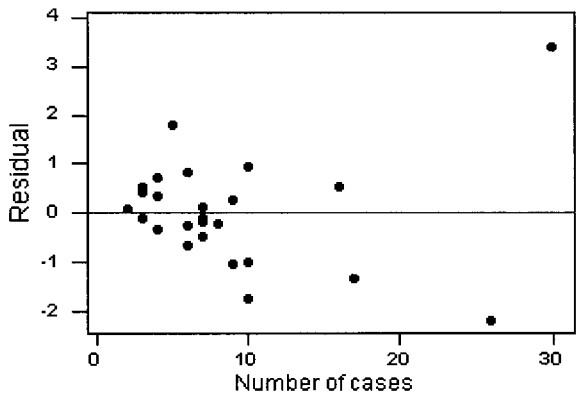


Figure 3.21: Residual plot: \hat{r}_i versus x (number of cases)

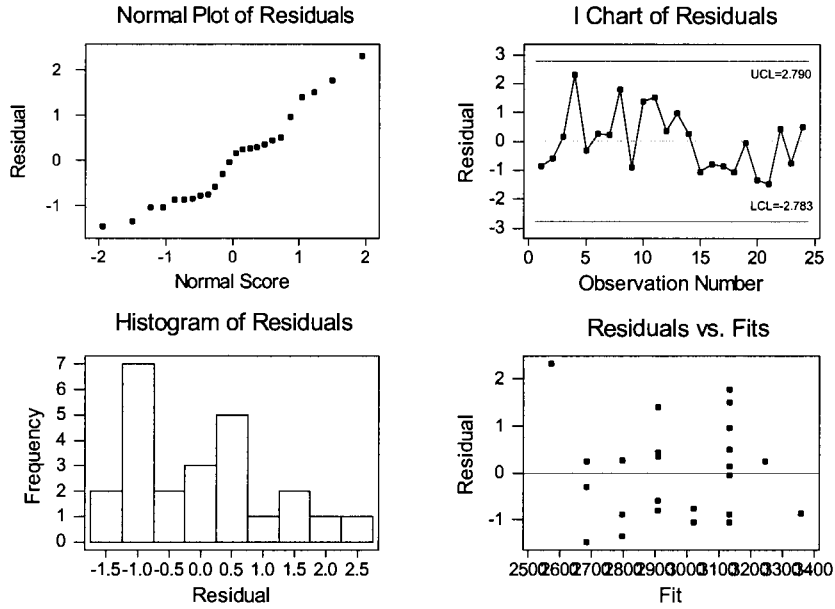


Figure 3.22: Variable plots of residuals \hat{r}_i for birth weight data

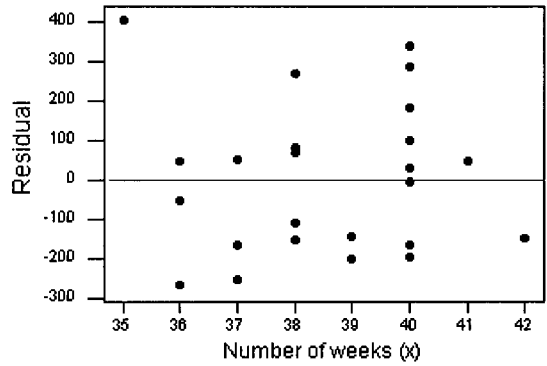


Figure 3.23: Residual plot: \hat{r}_i versus x (age)

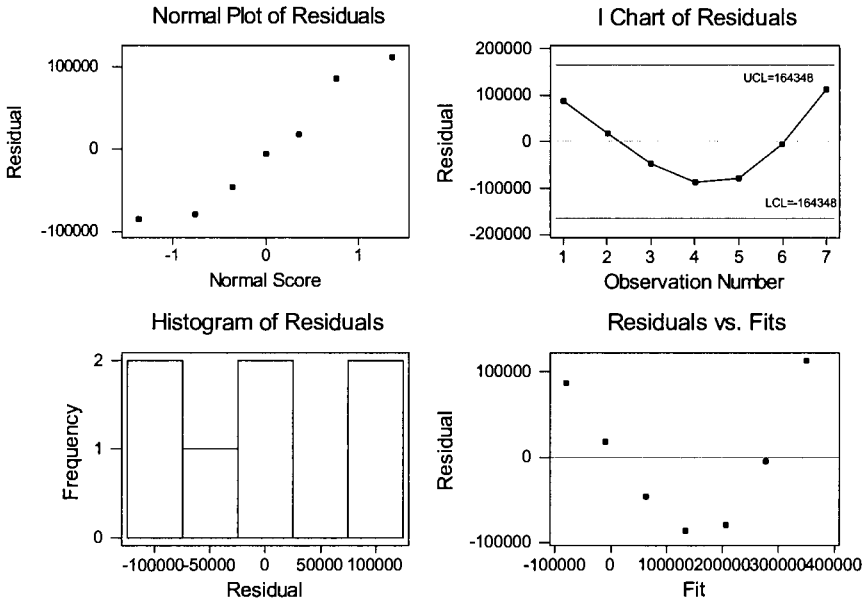


Figure 3.24: Variable plots of residuals \hat{r}_i for Clark county population data

3.10.1 Transformations of x

If the true model is of the form

$$Y_i = \beta_0 + \beta_1 g(x_i) + \varepsilon_i \quad (3.224)$$

where $g(x)$ is a known function of x , then letting $z_i = g(x_i)$ (3.224) takes the form

$$Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, 1 \leq i \leq n, \quad (3.225)$$

and we have a linear model in z . In this case if the errors are independent and $N(0, \sigma^2)$, the analysis for (3.225) can be carried out as before. The transformation $z = g(x)$ merely rescales the independent variable and has no effect on the linearity of the model, since in statistical terms linearity is determined by the linearity in the parameters and not that of the variable x .

The need for a possible transformation of the independent variable may be indicated by the plot of $\hat{\varepsilon}_i$ (or \hat{r}_i) versus x or curvature in the plot of $\hat{\varepsilon}_i$ versus \hat{y}_i and/or theoretical considerations; the latter being perhaps the best indicator as to the choice of the transformation $g(x)$.

Example 3.27 Some of the transformations of x for typical nonlinearities in x are \sqrt{X} , $\exp(X)$, and $\exp(-X)$.

Although theoretical considerations concerning the origin of the data can often be helpful in selecting a transformation, in most cases, if such a transformation is suggested,

there may not be any immediately obvious choice for g in (3.224). In such a situation several procedures have been suggested for letting the data guide the choice. For example, Box and Tidwell in [12] give a method for the determination of λ in a model of the form

$$Y = \beta_0 + \beta_1 x^\lambda + \varepsilon. \quad (3.226)$$

As this procedure requires the use of multiple regression techniques we will defer our analysis of this method to Chapter 6.

3.10.2 Transformations in x and y

If a simple transformation of the independent variable is insufficient to linearize the model, it may be necessary to investigate transformations of both x and y . In this regard there are a number of standard functional relations which can linearize the model using a variety of algebraic and/or analytic manipulations. Such nonlinear models are usually referred to as *intrinsically linear*.

As an example, consider the model

$$Y_x = \beta_0 e^{\beta_1 x} \varepsilon_x \quad (3.227)$$

where $\beta_0 > 0$ and ε_x is a lognormal random variable (i.e., $\log \varepsilon_x$ is normal). Then taking the logarithm of Y_x we get

$$\log Y_x = \log \beta_0 + \beta_1 x + \log \varepsilon_x, \quad (3.228)$$

which is a linear model in the transformed variables $Z_x = \log Y_x$, $\beta'_0 = \log \beta_0$, β_1 and $\varepsilon'_x = \log \varepsilon_x$. In this case, if the errors $\log \varepsilon_x$ are independent and $N(0, \sigma^2)$ then $\log \beta_0$ and β_1 may be estimated by least squares and the model analyzed using the techniques developed previously.

Example 3.28 Consider the exponential growth model, $Y_x = \exp(\beta_0 + \beta_1 x + \varepsilon_x)$, for the Clark County population data (in Example 3.20). Then the model is, in fact, intrinsically linear, hence it is linearizable. By taking the logarithm of y , as we have seen in Figure 3.13, the relationship between the size of the population and time (in years) shows almost linear growth.

One should note that in order for a model to be intrinsically linear, both the parameters and the errors must appear linearly in the transformed model. For instance, if

$$Y_x = \beta_0 e^{\beta_1 x} + \varepsilon_x \quad (3.229)$$

then taking the log of Y_x will not linearize (3.229), since $\log(\beta_0 e^{\beta_1 x} + \varepsilon_x) \neq \log(\beta_0 e^{\beta_1 x}) + \log \varepsilon_x$.

For convenient reference we give a list of commonly occurring intrinsically linear models and their equivalent linear forms in Table 3.24. In addition, graphs of the shapes of these functions are given in Figure 3.25(a)-(e). These may be helpful in identifying possible transformations to apply after examining a scatter plot of the data. Since

the graphs of different functions have similar shapes, keep in mind that more than one transformation may give an acceptable fit over a given range of data.

Table 3.24 Intrinsic Nonlinear Forms and Transformations

Figure	Prototype Form	Type of Transformation	Linearized Form
3.25(a)	$Y = \exp(\beta_0 + \beta_1 x)$	logarithm: $Y' = \log Y$	$Y' = \beta_0 + \beta_1 x$
3.25(b)	$Y = \beta_0 x^{\beta_1}$	logarithm: $\begin{cases} Y' = \log Y \\ x' = \log x \end{cases}$	$Y' = \log \beta_0 + \beta_1 x'$
3.25(c)	$Y = \beta_0 + \beta_1 \log x$	logarithm: $x' = \log x$	$Y' = \beta_0 + \beta_1 x'$
3.25(d)	$Y = x / (\beta_0 x - \beta_1)$	reciprocal (inverse) : $\begin{cases} Y' = 1/Y \\ x' = 1/x \end{cases}$	$Y' = \beta_0 - \beta_1 x'$
3.25(e)	$Y = 1 / (1 + e^{\beta_0 + \beta_1 x})$	reciprocal & logarithm: $Y' = \log(1/Y - 1)$	$Y' = \beta_0 + \beta_1 x$

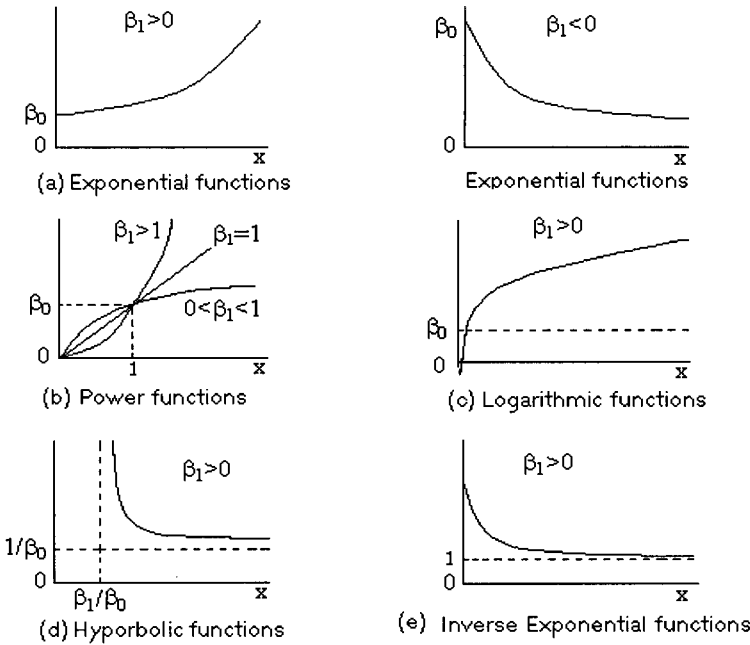


Figure 3.25: Curves of linearizable functions

3.10.3 Box-Cox Transformations

Since an appropriate transformation of the data may not always be immediately apparent from the examination of scatter plots, it is helpful to have procedures which will allow the data to help select the form of the transformation. In this section we will discuss a widely used procedure developed by Box and Cox [9] and generalized by a number of others in [6, 20].

In situations where the errors have a nonnormal distribution, such as in the logarithmic model (3.228), one generally would like to have a transformation of Y which not only linearizes the model but also transforms the errors to be approximately normal. A family of transformations, which includes the logarithm that has been found useful in this regard is the *power family*

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1) / \lambda, & \lambda \neq 0, \\ \log y, & \lambda = 0, \end{cases} \quad (3.230)$$

where we assume that our original data y is positive. (If not, then one can add a positive constant to each observation and analyze the resulting transformed data.) One should note that $\log y$ is the limiting value of $y^{(\lambda)}$ as $\lambda \rightarrow 0$. That is,

$$\lim_{\lambda \rightarrow 0} \frac{y^\lambda - 1}{\lambda} = \log y. \quad (3.231)$$

Suppose now that our data $Y_i, 1 \leq i \leq n$, are such that $Y_i^{(\lambda)}, 1 \leq i \leq n$, are independent and $N(0, \sigma^2)$ and

$$E[Y_i^{(\lambda)}] = \beta_0 + \beta_1 x_i, \quad (3.232a)$$

so that $Y_i^{(\lambda)}$ satisfy the conditions for the simple linear regression model. Since the value of λ is generally unknown, our problem will be to simultaneously estimate all of the parameters $(\lambda, \beta_0, \beta_1, \sigma)$. Although a number of procedures have been proposed for doing this [27, 20, 6], maximum likelihood estimation appears to be the most popular. This is the approach we will follow.

If we let

$$\mu_i = \beta_0 + \beta_1 x_i \quad (3.233)$$

then it follows from the change of variables formula [40] that the density of Y_i is given by

$$f_{Y_i}(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} [y_i^{(\lambda)} - \mu_i]^2 \right\} y_i^{\lambda-1} \quad (3.234)$$

where

$$y_i^{(\lambda)} = \begin{cases} (y_i^\lambda - 1) / \lambda, & \lambda \neq 0, \\ \log y_i, & \lambda = 0. \end{cases} \quad (3.235)$$

The likelihood function L for the n observations (y_1, y_2, \dots, y_n) is then given by

$$L = \prod_{i=1}^n f_{Y_i}(y_i) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i^{(\lambda)} - \mu_i]^2 \right\} J(\lambda) \quad (3.236)$$

where $J(\lambda) = \prod_{i=1}^n y_i^{\lambda-1} = (\prod_{i=1}^n y_i)^{\lambda-1}$. The log likelihood $\mathcal{L} = \log L$ is then given by

$$\mathcal{L} = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i^{(\lambda)} - \mu_i]^2 + \log J(\lambda). \quad (3.237)$$

To find the MLEs of $(\lambda, \beta_0, \beta_1, \sigma)$ we differentiate \mathcal{L} with respect to these parameters and then solve the resulting equations obtained by setting these derivatives equal to zero.

Unfortunately, these equations do not have an explicit analytical solution, so we will have to resort to numerical techniques for solving them.

Differentiating (3.237) with respect to (β_0, β_1) it follows easily that the MLE estimates of (β_0, β_1) (denoted by $\hat{\beta}_0(\lambda), \hat{\beta}_1(\lambda)$) are given by finding the least squares estimates of (β_0, β_1) using the transformed data $y_i^{(\lambda)}$. Thus, if λ is known, the remaining analysis can be carried out using the usual least squares theory. The remaining problem is to estimate σ and λ .

Differentiating (3.237) with respect to σ shows that the maximum likelihood estimate $\hat{\sigma}(\lambda)$ of σ is given by

$$\hat{\sigma}(\lambda) = \left\{ \frac{1}{n} \sum_{i=1}^n \left[y_i^{(\lambda)} - \hat{y}_i^{(\lambda)} \right]^2 \right\}^{1/2} \quad (3.238)$$

where $\hat{y}_i^{(\lambda)} = \hat{\beta}_0(\lambda) + \hat{\beta}_1(\lambda) x_i$. Thus the value of \mathcal{L} maximized with respect to $(\beta_0, \beta_1, \sigma)$ is given by

$$\mathcal{L}_{\max} = -\frac{n}{2} \log 2\pi - n \log \hat{\sigma}(\lambda) + \log J(\lambda) - \frac{n}{2}. \quad (3.239)$$

To estimate λ we must maximize \mathcal{L}_{\max} with respect to λ . Since the constant $-(n/2) \log 2\pi - n/2$ is unimportant in this respect, it suffices to maximize

$$\mathcal{L}' = -n \log \hat{\sigma}(\lambda) + \log J(\lambda) \quad (3.240)$$

As \mathcal{L}' is a complicated function of λ some type of numerical procedure is needed to do this. One method, suggested in [27], is to start with a plausible range of λ , say $[-2, 2]$, evaluate \mathcal{L}' at a set of values in $[-2, 2]$, make a smooth graph from the resulting points, and then select the value $\hat{\lambda}$ which maximizes \mathcal{L}' by inspection. Draper and Smith in [27] state that 10-20 evenly spaced points is usually adequate.

Having done this, one can then obtain an approximate confidence interval for λ and test the hypothesis $H_0: \lambda = 1$ for the need of a transformation. If H_0 is rejected, then we transform the data using $\hat{\lambda}$ and the MLE of the parameters are

$$\left\{ \hat{\lambda}, \hat{\beta}_0(\hat{\lambda}), \hat{\beta}_1(\hat{\lambda}), \hat{\sigma}(\hat{\lambda}) \right\}. \quad (3.241)$$

This test can be performed using the approximate confidence interval for λ , which is based on the difference between the two likelihood ratios, $\mathcal{L}_{\max}(\hat{\lambda}) - \mathcal{L}_{\max}(\lambda_0)$, [27, 5]. Hence, an approximate $(1 - \alpha)$ 100% confidence interval for λ consists of those values of λ_0 which satisfy the inequality

$$-\frac{n}{2} \log \hat{\sigma}^2(\hat{\lambda}) - \left[-\frac{n}{2} \log \hat{\sigma}^2(\lambda_0) \right] \leq \frac{1}{2} \chi_{\alpha}^2(1) \quad (3.242)$$

where $\chi_{\alpha}^2(1)$ is the upper α -th percent point of the χ^2 -distribution with one degree of freedom.

Some authors recommend rounding $\hat{\lambda}$ to the nearest 1/4 or 1/3 for ease of interpretation. For example, if $\hat{\lambda} = 1.43$ then the value $\hat{\lambda} = 1.5$ would be used in (3.240). This procedure seems reasonable and we would recommend it.

Before presenting a numerical example of this procedure let us point out that standard regression analysis software may not directly accommodate this method due to the

necessity of having to compute \mathcal{L}' . However, a slight modification of the method enables one to use standard regression software for estimating the parameters.

For this, observe that in (3.240)

$$-n \log \hat{\sigma}(\lambda) + \log J(\lambda) = \log \left\{ \left[\frac{J^{1/n}(\lambda)}{\hat{\sigma}(\lambda)} \right]^n \right\}. \quad (3.243)$$

Thus to maximize \mathcal{L}' it suffices to minimize

$$s(\lambda) = \frac{\hat{\sigma}(\lambda)}{[J(\lambda)]^{1/n}} = \frac{\hat{\sigma}(\lambda)}{(\dot{y})^{\lambda-1}} \quad (3.244)$$

where $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$ is the *geometric mean* of the observations. Now a little algebra shows that $ns^2(\lambda)$ is the sum of squares of the residuals obtained by regressing

$$\frac{y^{(\lambda)}}{(\dot{y})^{\lambda-1}} = \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} \quad (3.245)$$

on x . Using this observation we obtain the following algorithm for fitting (3.240).

- (i) Choose a range of λ , $I = [\lambda_{\min}, \lambda_{\max}]$ and points $\lambda \in I$ to evaluate $s(\lambda)$ as before.
- (ii) Regress $y^{(\lambda)} / (\dot{y})^{\lambda-1}$ on x and obtain $ns^2(\lambda)$, the sum of squares of the residuals, from these regressions.
- (iii) Plot the values $(\lambda, ns^2(\lambda))$ and obtain $\hat{\lambda}$ as before.
- (iv) With the value of $\hat{\lambda}$ chosen above, regress $y^{(\hat{\lambda})}$ on x and continue the analysis on the linearized model.

Example 3.29 (Clark county population data) Recall the Clark county population data in Example 3.20. We selected values of λ , ranging from -1.0 to 1.0 , and for each chosen λ the transformation (3.230) was made and the linear regression of $y^{(\lambda)}$ on x was obtained.

Note from Table 3.25 that the Box-Cox procedure identifies a power near $\lambda = 0$ as being the optimum value.

Table 3.25 Values of SSE for Selected Values of λ

λ	SSE	λ	SSE
-1.0	34,951,428,737	0.1	5,144,162
-0.8	10,208,672,414	0.2	37,404,370
-0.6	2,505,000,472	0.4	461,173,312
-0.4	437,511,496	0.6	2,625,347,048
-0.2	35,247,919	0.8	10,652,489,336
-0.1	4,898,279	1.0	36,308,540,498

The approximate 95% confidence interval for λ is specified on the graph ($-0.577 \leq \lambda \leq 0.481$) in Figure 3.26. We see that the validity of using $\lambda = 0$ (i.e., $\log Y$) is confirmed by this graph and that the transformation is well estimated.

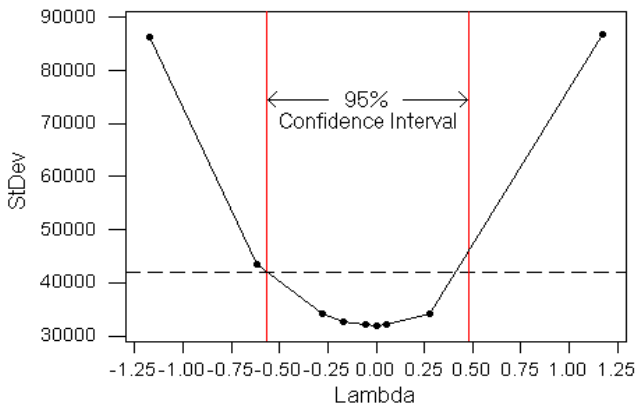


Figure 3.26: An Approximate 95% Confidence Interval for λ

Application of the natural logarithm transformation to the original data gives us the transformed data in Table 3.26.

Table 3.26 Transformed Values $y^{(\lambda)} = \log Y$			
Obs. No.	Year (x)	Population (Y)	$\log_e Y$
1	1920	4,859	8.4886
2	1930	8,539	9.0524
3	1940	16,414	9.7059
4	1950	48,589	10.7912
5	1960	127,016	11.7521
6	1970	273,288	12.5183
7	1980	463,087	13.0457

3.11 Exercises

- 3.1** Suppose we have the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$. Define the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$. Show that

(a) $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

(b) $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$.

(c) $\sum_{i=1}^n x_i \hat{\varepsilon}_i = 0$.

(d) $\sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i = 0$.
- 3.2** A researcher considered a simple linear model for analyzing a certain data set with $n = 26$ pairs of observations. From calculations, he obtained estimates: $\hat{\beta}_0 = 8.0$, $\hat{\beta}_1 = 2.2$ and $\hat{\sigma}^2 = 9$.

(a) If $\bar{x} = 7$, what is the value of \bar{y} ?

(b) Find the (approximate) distribution of Y_i specifying $E(Y_i)$ and $Var(Y_i)$.

(c) Suppose that an observation on Y is made for $x = 5$. Find the probability that Y falls between 16 and 22.

3.3 Suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least square estimators of the simple linear regression model parameters. Find $Cov(\hat{\beta}_0, \hat{\beta}_1)$.

3.4 Verify the expression in (3.29). [Hint: Use the Pythagorean Theorem.]

3.5 Find the maximum likelihood estimator of σ^2 for the error model in (3.4).

3.6 Given the following partial calculations which were made from a sample of size $n = 42$:

$$\bar{x} = 6.1, S_{xx} = 9.74, \hat{\beta}_0 = 3.45, \hat{\beta}_1 = 1.81, \hat{\sigma}^2 = 4.3.$$

(a) Construct 90% confidence intervals for β_0 , for β_1 , and σ^2 .

(b) Test $H_0 : \beta_1 = 0$ at $\alpha = 5\%$.

(c) Obtain a 90% confidence region for β_0 and β_1 .

3.7 For the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$, the sum of squares due to regression is defined by

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

(a) Show that $\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$.

(b) Use (a) to show that $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 S_{xx}$.

(c) Use (b) to establish the relationship

$$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}.$$

(c) Show that $E(MSR) = \sigma^2 + \beta_1^2 S_{xx}$.

3.8 A group of researchers studied the effect of the molar ratio of sebacic acid on the intrinsic viscosity of copolyesters. The following table provides the data.

Molar Ratio (x)	Viscosity (y)
0.3	0.44
0.4	0.55
0.5	0.57
0.6	0.70
0.7	0.58
0.8	0.34
0.9	0.20
1.0	0.45

(a) Draw a scatterplot for these data.

(b) Fit the simple linear regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

(c) Calculate the residuals, $\hat{\varepsilon}_i$ and plot them.

(d) Is the model considered in (b) significant? Use $\alpha = 0.05$.

- 3.9** Observations on the yield of a chemical reaction taken at various levels of temperatures were recorded as follows [90].

Temperature x ($^{\circ}C$)	Yield y (%)		
150	77.4	76.7	78.2
200	84.1	84.5	83.7
250	88.9	89.2	89.7
300	94.8	95.9	94.7

Suppose that a simple linear regression model was postulated.

- Draw a scatterplot for this data.
- Find the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Estimate the mean square error (MSE) $\hat{\sigma}^2$.
- Find the standard error of $\hat{\beta}_1, \hat{\sigma}(\hat{\beta}_1)$, and perform the t -test for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$. Take $\alpha = 5\%$.
- Construct a 95% confidence interval for $\hat{\beta}_1$.
- Construct a 90% prediction interval for the mean response at $x = 275^{\circ}C$.

- 3.10** Consider the data given in Exercise 3.9.

- Construct the ANOVA table with lack of fit and pure error terms.
- Is the regression model significant in (a)? Take $\alpha = 0.01$.
- Show numerically that $F = t^2$, where t was in Exercise 3.9(d) for testing $\beta_1 = 0$.
- Test the lack of fit of the model. Use $\alpha = 0.05$.
- Compute R^2 and give an interpretation.

- 3.11** Consider a simple linear regression model with $\beta_0 = 0$.

- Verify that the least-squares estimator of $\beta_1, \hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

- Show that $\hat{\beta}_1$ is unbiased and that $Var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n x_i^2$.
- Show that s^2 in (3.164) is unbiased for σ^2 . How many degrees of freedom does it have?
- Show that $SSR = \sum_{i=1}^n (\hat{\beta}_1 x_i)^2$.

- 3.12** The data below for the years 1919 to 1935 ($n = 17$), gives x = the water content of snow on April 1 and Y = the water yield from April to July (in inches) in the

Snake River watershed in Wyoming [121].

Year	x	Y	Year	x	Y
1919	23.1	10.5	1928	37.9	22.8
1920	32.8	16.7	1929	30.5	14.1
1921	31.8	18.2	1930	25.1	12.9
1922	32.0	17.0	1931	12.4	8.8
1923	30.4	16.3	1932	35.1	17.4
1924	24.0	10.5	1933	31.5	14.9
1925	39.5	23.1	1934	21.1	10.5
1926	24.3	12.4	1935	27.6	16.1
1927	52.5	24.9			

- Fit a regression through the origin with the model $Y_i = \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n$.
- Estimate σ^2 .
- Construct a 95% confidence interval for β_1 .
- Obtain the ANOVA table and compute F .

3.13 After the values of Y were regressed on x , the following numbers were reported:

$$\hat{\beta}_0 = 12.5, \hat{\beta}_1 = -5.73, \bar{y} = 28.8.$$

- Complete the ANOVA table:

Source of Variation	df	Sum of Squares	Mean Squares
Regression		157.94	
Residual	19		
Total (Corrected)		188.25	

- Calculate the proportion of the Y -variability that is explained by the fitted regression line.
- Find the sample correlation coefficient between x and y .
- Obtain a 95% confidence interval for β_1 . [Hint: $\hat{\beta}_1^2 = SSR / \sum_{i=1}^n (x_i - \bar{x})^2$.]
- Test the hypothesis $H_0 : \beta_0 = 10$ against $H_1 : \beta_0 > 10$. Take $\alpha = 5\%$.

3.14 Consider a simple linear regression model in (3.4). Suppose we replace each x_i with cx_i , where c is a nonzero constant. That is, we have the model

$$Y_i = \beta_0 + \beta_1 (cx_i) + \varepsilon_i, i = 1, 2, \dots, n.$$

- Derive the least squares estimators of β_0 and β_1 .
- Estimate σ^2 .
- Are these estimators in (a) the same as (3.18)-(3.19)?

- 3.16** Show that the estimator (3.19), $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ can be expressed as $\hat{\beta}_0 = \sum_{i=1}^n c_i y_i$, where

$$c_i = \frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{S_{xx}}.$$

- 3.17** Consider a simple linear regression model with the assumption that ε_i 's are i.i.d. $N(0, \sigma^2)$. Suppose we reparameterize the model as

$$Y_i = \gamma_0 + \gamma_1 (x_i - \bar{x}) + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the MLEs of β_0 and β_1 , respectively and $\hat{\gamma}_0$ and $\hat{\gamma}_1$ denote the MLEs of γ_0 and γ_1 , respectively.

- (a) Show that $\hat{\gamma}_1 = \hat{\beta}_1$.
 (b) Show $\hat{\gamma}_0 \neq \hat{\beta}_0$. In fact, show that $\hat{\gamma}_0 = \bar{y}$.
 (c) Find the distribution of $\hat{\gamma}_0$.
- 3.18** Prove that two expressions for the sum of squares due to regression are equal. That is, show that

$$\sum_{i=1}^n (\hat{y} - \bar{y})^2 = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- 3.19** After observations (x_i, y_i) were obtained, we postulated the regression model to describe the relationship by

$$Y_i = \theta_1 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

where the ε_i 's are assumed to be i.i.d. $N(0, \sigma^2)$. This is a quadratic model passing through the origin.

- (a) Find the least squares estimator of θ_1 .
 (b) Find the maximum likelihood estimator of θ_1 .
- 3.20** The *minimum absolute deviation line* is given by the values of δ_0 and δ_1 that minimize

$$\sum_{i=1}^n |y_i - (\delta_0 + \delta_1 x_i)|.$$

(a) Show that, for a data set with three observations, (x_1, y_1) , (x_1, y_2) , and (x_3, y_3) , any line that goes through (x_3, y_3) and lies between (x_1, y_1) and (x_1, y_2) is a minimum absolute deviation line.

Suppose we now have three pairs of measurements that are taken on a heart rate (x , in beats per minute) and oxygen consumption (y , in ml/kg); (127, 14.4), (127, 11.9), and (136, 17.9).

- (b) Find the slope and intercept of the least squares line.
 (c) Find the minimum absolute deviation line.

- 3.21** Given the data of pairs (x, y) tabulated below, and assuming normal, independent observations with constant variance σ^2 :

y	-2.4	5.3	-0.6	8.4	0.4	-1.0	8.7	-4.6
x	-0.2	2.2	-0.1	2.8	1.5	2.1	2.9	-0.5

- (a) Find the maximum likelihood estimates of β_0 , β_1 , and σ^2 .
- (b) Do the MLEs in (a), $\tilde{\beta}_0$ and $\tilde{\beta}_1$, agree with the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$?
- (c) Construct a 90% confidence interval for the mean response of Y_x when $x = 0.5$.
- (d) Make a normal probability plot for the residuals.
- 3.22** Consider the model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i 's are i.i.d. $N(0, k_i^{-1} \sigma^2)$ and k_i 's are known positive constants, $i = 1, 2, \dots, n$.
- (a) Write the log-likelihood function, $\mathcal{L}(\beta_0, \beta_1, \sigma^2)$.
- (b) Find the MLEs of β_0 and β_1 .
- (c) Derive the normal equations and find the LSEs of β_0 and β_1 .

Chapter 4

Random Vectors and Matrix Algebra

4.1 Introduction

In dealing with the simple linear regression model we were able to calculate all of the necessary quantities without using any special algebraic techniques. However, to extend regression models to account for more than one independent variable it becomes convenient to use the more sophisticated approach afforded by matrix algebra.

As a consequence, this chapter will be devoted to developing some of the basic concepts in this area and to proving a number of important theorems which the reader may not have encountered in elementary courses in matrix algebra. In addition, some probabilistic and statistical consequences of these ideas will also be examined, particularly the joint multivariate normal distribution. Because our statistical work in future chapters will be restricted almost exclusively to full rank regression models (see Chapter 5), some specialized topics such as the theory of generalized inverses will be omitted. Further algebraic results of this and a more advanced nature can be found in [112] and we refer the reader there for more details.

4.2 Matrices and Vectors

A *matrix* is a rectangular array of numbers as depicted in Figure 4.1. The horizontal sequences are called the *rows* of the matrix, while the vertical sequences are called its *columns*. The rows are labeled in increasing order from top to bottom, while the columns are labeled from left to right.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & \cdot & a_{1m} \\ a_{21} & \cdot & \cdots & \cdot & a_{2m} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & a_{ij} & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ a_{n1} & \cdot & \cdots & \cdot & a_{nm} \end{bmatrix}$$

Figure 4.1: $n \times m$ Matrix

A matrix with n rows and m columns, $n \geq 1$, $m \geq 1$, will be called an $n \times m$ *matrix* (“ n by m ” matrix) and the pair of integers (n, m) its *dimension*. The individual numbers that make up the matrix are usually referred to as its *elements* and the number in the i -th row and j -th column will be called the ij -th element. In general, a matrix will be denoted by an *upper case boldface letter* \mathbf{A} , \mathbf{B} , \mathbf{X} etc., and the corresponding elements by lower case letters a_{ij} , b_{ij} , etc. If $n = m$, the matrix is *square*, otherwise it is *rectangular*.

If \mathbf{A} is an $n \times n$ square matrix, then the elements $a_{ii} \equiv a_i$, $1 \leq i \leq n$, are called the *diagonal elements* of \mathbf{A} , while if $i \neq j$, a_{ij} are called the *off-diagonal elements*.

4.2.1 Some Special Matrices

In using matrix algebra to analyze regression models a number of special matrixes occur repeatedly so it is convenient to have special notation reserved for these circumstances. If the matrix \mathbf{A} has all its elements zero, then it will be called a *zero* matrix and denoted by $\mathbf{0}$. If the dimension is not clear from the context, then $\mathbf{0}_{n \times m}$ will denote then $n \times m$ zero matrix.

A square matrix \mathbf{A} whose off-diagonal elements are all zero will be called a *diagonal matrix* and often written as $\text{diag}(a_i)$ where a_i , $1 \leq i \leq n$, are the diagonal elements. If $a_i = 1$, $1 \leq i \leq n$, then \mathbf{A} is the $n \times n$ *identity matrix* written as \mathbf{I}_n or just \mathbf{I} if the dimension is understood. When all the elements of a square matrix above the diagonal are zero, then the matrix is said to be *lower triangular*, while if all the elements below the diagonal are zero, then the matrix is said to be *upper triangular*.

Given a matrix \mathbf{A} we can form a new matrix \mathbf{A}^T called the *transpose* of \mathbf{A} by interchanging the rows and columns of \mathbf{A} . that is, the i -th column \mathbf{A}^T is the i -th row of \mathbf{A} . For example, if

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix},$$

then

$$\mathbf{A}^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

If \mathbf{A} is $n \times m$, then \mathbf{A}^T is $m \times n$ and \mathbf{A}^T has the important property that if you transpose it, then you arrive at the original matrix \mathbf{A} . Symbolically, $(\mathbf{A}^T)^T = \mathbf{A}$. If \mathbf{A} is square and $\mathbf{A} = \mathbf{A}^T$, then we say that \mathbf{A} is *symmetric*, otherwise it is *non-symmetric*. For example,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$$

is symmetric, while

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

is not.

As we shall see, symmetric matrices play an important role in regression analysis, and a number of their properties will be studied further in Section 4.6.

4.3 Fundamentals of Matrix Algebra

As for real numbers, we can introduce various operations on matrices, such as addition, subtraction and multiplication. This topic is usually referred to as *matrix algebra* and we will take up some basic aspects in this section. Further details will be developed in later sections.

4.3.1 Matrix Addition

If \mathbf{A} and \mathbf{B} are two $n \times m$ matrices, they can be added together to give a new matrix \mathbf{C} defined by the formula $\mathbf{C} = [c_{ij}]$, $1 \leq i \leq n$, $1 \leq j \leq m$, where

$$c_{ij} = a_{ij} + b_{ij}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m. \quad (4.1)$$

\mathbf{C} is called the *sum* of \mathbf{A} and \mathbf{B} and is written as

$$\mathbf{C} = \mathbf{A} + \mathbf{B}. \quad (4.2)$$

Note that by definition only matrices of the same dimension can be added and that the sum of two $n \times m$ matrices is again an $n \times m$ matrix, in which case they are said to be *conformable* for addition. Matrix addition has many of the properties of addition of real numbers, and the corresponding names are the same. A number of these are stated below and their proofs are left as simple exercises.

4.3.2 Properties of Matrix Addition

If \mathbf{A} , \mathbf{B} , \mathbf{C} , etc. are $n \times m$ matrices, then the following properties hold:

- (i) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ (commutativity)
- (ii) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$ (associativity)
- (iii) From (ii) and (iii) it follows that a sequence of n matrices can be added without regard to order and with no need for parentheses. For example, we have for any $n \times m$ matrices \mathbf{A} , \mathbf{B} , \mathbf{C} , \mathbf{D} that

$$(\mathbf{A} + \mathbf{B}) + (\mathbf{C} + \mathbf{D}) = \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D}, \text{ etc.} \quad (4.3)$$

More generally, the sum of the $n \times m$ matrices \mathbf{A}_i , $1 \leq i \leq p$, is defined in any order and this allows us to use summation notation unambiguously. That is,

$$\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_p \equiv \sum_{i=1}^p \mathbf{A}_i. \quad (4.4)$$

- (iv) If \mathbf{A} is $n \times m$ and $\mathbf{0}$ is the $n \times m$ zero matrix, then we have

$$\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}. \quad (4.5)$$

- (v) For any matrix \mathbf{A} we can define its negative $-\mathbf{A}$, given by $-\mathbf{A} = [-a_{ij}]$, and from this it follows easily that

$$\mathbf{A} + (-\mathbf{A}) = -\mathbf{A} + \mathbf{A} = \mathbf{0}. \quad (4.6)$$

If we introduce the concept of matrix subtraction given by

$$\mathbf{A} - \mathbf{B} \equiv \mathbf{A} + (-\mathbf{B}) \quad (4.7)$$

where \mathbf{A} and \mathbf{B} are $n \times m$, then (4.7) can be written more conveniently as

$$\mathbf{A} - \mathbf{A} = \mathbf{0} \quad (4.8)$$

in analogy with the corresponding property of real numbers.

4.3.3 Matrix Multiplication

There are a number of ways of defining matrix multiplication with pointwise multiplication of matrices perhaps being the most natural. Unfortunately, that definition is not the one that has been found to be most useful in practice and so a different, somewhat more complicated method, is most often used.

The definition of matrix multiplication is perhaps most easily motivated by trying to find a compact notation to express a system of n linear equations in n unknowns. For instance, consider the system of equations

$$\begin{cases} 2x + 3y = 5 \\ 5x + 6y = 7 \end{cases} \quad (4.9)$$

and let \mathbf{A} be the *coefficient matrix* of (4.9). That is

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}, \quad (4.10)$$

and let $\mathbf{x} = (x, y)^T$ be the column vector of unknowns. Then we can define the *product* of \mathbf{A} and \mathbf{x} so that the result is the left-hand side (4.10). That is,

$$\mathbf{Ax} = \begin{bmatrix} 2x + 3y \\ 5x + 6y \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (4.11)$$

From (4.11) we see that in order to get the first row of \mathbf{Ax} we multiply the first element in the first row of \mathbf{A} by the first element of \mathbf{x} , then we multiply the second element in the first row of \mathbf{A} by the second element of \mathbf{x} and the results are added. The second row of \mathbf{Ax} is produced in a similar fashion.

More generally, if \mathbf{A} is $n \times m$ and \mathbf{B} is $m \times p$ then we can define the *product* of \mathbf{A} and \mathbf{B} , denoted by \mathbf{AB} , by (the 'bars' separate columns of \mathbf{B})

$$\mathbf{AB} = \mathbf{A} [\mathbf{b}_1 | \mathbf{b}_2 | \cdots | \mathbf{b}_p] = [\mathbf{Ab}_1 | \mathbf{Ab}_2 | \cdots | \mathbf{Ab}_p] \quad (4.12)$$

where \mathbf{b}_j , $1 \leq j \leq p$, is the j -th column of \mathbf{B} , and \mathbf{Ab}_j is the product of \mathbf{A} with the column vector \mathbf{b}_j . More explicitly, if $\mathbf{b}_j = (b_{1j}, b_{2j}, \dots, b_{mj})^T$ then the i -th element $(\mathbf{Ab}_j)_i$ of \mathbf{Ab}_j is given by

$$(\mathbf{Ab}_j)_i = \sum_{k=1}^m a_{ik} b_{kj} \quad (4.13)$$

From (4.12)-(4.13) we see that if $\mathbf{C} = \mathbf{AB}$ is the product of \mathbf{A} and \mathbf{B} , then,

$$c_{ij} = \sum_{k=1}^m a_{ik}b_{kj}, \quad 1 \leq i \leq n, 1 \leq j \leq p. \quad (4.14)$$

In particular, it is important to note that \mathbf{AB} is defined if and only if the number of columns of \mathbf{A} equals the number of rows of \mathbf{B} . Thus if \mathbf{A} is $n \times m$ and \mathbf{B} is $m \times p$, then \mathbf{AB} is well defined and \mathbf{AB} is $n \times p$. In this case we say that \mathbf{A} and \mathbf{B} are *conformable* matrices. Thus, in general, if \mathbf{AB} is defined then \mathbf{BA} is not defined unless $p = n$. In particular, if \mathbf{A} and \mathbf{B} are both square $n \times n$ matrices, then both \mathbf{AB} and \mathbf{BA} are defined but in general are unequal. If it is true that $\mathbf{AB} = \mathbf{BA}$ we then say that \mathbf{A} and \mathbf{B} *commute*.

4.3.4 Properties of Matrix Multiplication

As for matrix addition, matrix multiplication satisfies a number of useful properties which the student should commit to memory. The details are left as exercises.

(i) In general, $\mathbf{AB} \neq \mathbf{BA}$.

(ii) If, however, \mathbf{A} is $n \times m$, \mathbf{B} is $m \times p$ and \mathbf{C} is $p \times q$ then

$$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} = \mathbf{ABC}. \quad (4.15)$$

Equation (4.15) is called the *associative property* of matrix multiplication and is true for any sequence of p matrices $(\mathbf{A}_i)_{i=1}^p$ whose product $\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_p$ is well defined.

(iii) If \mathbf{I}_n is the $n \times n$ identity matrix and \mathbf{A} is $n \times m$, then

$$\mathbf{I}_n\mathbf{A} = \mathbf{A} \quad (4.16)$$

and similarly,

$$\mathbf{A}\mathbf{I}_m = \mathbf{A}. \quad (4.17)$$

In particular, if \mathbf{A} is $n \times n$ then

$$\mathbf{A}\mathbf{I}_n = \mathbf{I}_n\mathbf{A} = \mathbf{A}. \quad (4.18)$$

(iv) If $\mathbf{0}_{n \times m}$ is the $n \times m$ zero matrix and \mathbf{A} is $m \times p$ then

$$\mathbf{0}_{n \times m}\mathbf{A} = \mathbf{0}_{n \times p} \quad (4.19)$$

while

$$\mathbf{A}\mathbf{0}_{n \times p} = \mathbf{0}_{n \times m}. \quad (4.20)$$

(v) When \mathbf{AB} is defined, then

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \quad (4.21)$$

(vi) If \mathbf{A} is $n \times m$ and \mathbf{B} and \mathbf{C} are $m \times p$, then

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}. \quad (4.22)$$

Equation (4.22) is called the *distributive law* for matrix multiplication.

4.3.5 Scalar Multiplication

In addition to being able to multiply two matrices together it is also possible to multiply a matrix by a real number. This form of multiplication is usually called *scalar multiplication* and is defined as follows. If \mathbf{A} is an $n \times m$ matrix and c is a real number, then the scalar multiple of \mathbf{A} by c is the matrix $c\mathbf{A}$ whose ij -th element is just ca_{ij} . Thus $c\mathbf{A}$ is obtained by multiplying every element of \mathbf{A} by the real number c . This type of multiplication also has a number of useful properties which are stated below.

- (i) $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$;
- (ii) $(c + d)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$;
- (iii) $(cd)\mathbf{A} = c(d\mathbf{A})$;
- (iv) $0\mathbf{A} = \mathbf{0}$;
- (v) $1\mathbf{A} = \mathbf{A}$ and $(-1)\mathbf{A} = -\mathbf{A}$;
- (vi) $(c\mathbf{A})^T = c\mathbf{A}^T$;

Again (i)-(vi) are easily proved and are left as exercises.

4.3.6 Powers of Matrices

As for real numbers, it follows from the associative property of matrix multiplication that one can define powers of a square matrix \mathbf{A} .

If \mathbf{A} is an $n \times n$ matrix, then we can define $\mathbf{A}^1 \equiv \mathbf{A}$ and

$$\mathbf{A}^2 \equiv \mathbf{A}^1 \times \mathbf{A}^1 = \mathbf{A} \times \mathbf{A}. \quad (4.23)$$

Continuing this way

$$\mathbf{A}^3 = \mathbf{A}^2 \times \mathbf{A} = \mathbf{A} \times \mathbf{A}^2 \quad (4.24)$$

and by induction

$$\mathbf{A}^n = \mathbf{A}^{n-1} \times \mathbf{A} = \mathbf{A} \times \mathbf{A}^{n-1}. \quad (4.25)$$

\mathbf{A}^2 is read “A squared”, \mathbf{A}^3 is “A cubed” and \mathbf{A}^n is “A to the n -th power.” For some purposes it is useful to let

$$\mathbf{A}^0 \equiv \mathbf{I}_n. \quad (4.26)$$

From (4.25) and the associativity of matrix multiplication it follows that

$$\mathbf{A}^n \mathbf{A}^m = \mathbf{A}^m \times \mathbf{A}^n = \mathbf{A}^{n+m}, \quad n \geq 0, m \geq 0. \quad (4.27)$$

An important class of matrices in statistics (and elsewhere) are those that satisfy

$$\mathbf{A}^2 = \mathbf{A}. \quad (4.28)$$

When (4.28) holds we say that \mathbf{A} is an *idempotent* or *projection matrix*. In particular, \mathbf{I}_n is idempotent, as are diagonal $n \times n$ matrices of the form

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & 0 & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}_{n \times n} \quad (4.29)$$

with p ones and $n - p$ zeros on the diagonal and zeros elsewhere.

In addition, if \mathbf{A} is symmetric we will say that \mathbf{A} is an *orthogonal projection*. It will be shown in Section 4.6 that in an appropriate sense all orthogonal projection matrices are of the form in (4.29). This fact plays an important role in understanding the properties of linear regression models with normal errors.

4.3.7 Matrix Trace

If \mathbf{A} is an $n \times n$ matrix, then the *trace* of \mathbf{A} , $\text{tr}(\mathbf{A})$ is the sum of its diagonal elements; i.e.

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}. \quad (4.30)$$

$\text{Tr}(\mathbf{A})$ plays an important role in many regression calculations and has a number of important properties.

- (i) If $\mathbf{A} = \mathbf{0}$, then $\text{tr}(\mathbf{A}) = 0$.
- (ii) If $\mathbf{A} = \mathbf{I}_{n \times n}$, then $\text{tr}(\mathbf{A}) = n$.
- (iii) If $c \in \mathbb{R}$ then, $\text{tr}(c\mathbf{A}) = c \text{tr}(\mathbf{A})$.
- (iv) If \mathbf{A} and \mathbf{B} are $n \times n$ matrices, then

$$\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}). \quad (4.31)$$

- (v) If $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are $n \times n$ matrices, then

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}) \quad (4.32)$$

and

$$\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA}). \quad (4.33)$$

- (vi) $\text{tr}(\mathbf{A}^T) = \text{tr}(\mathbf{A})$.

The proofs of (i) to (vi) are left as exercises.

4.4 Matrices and Linear Transformations

Another way of regarding a matrix \mathbf{A} is to see how it acts on vectors by matrix multiplication. If \mathbf{x} is a column vector of dimension p and \mathbf{A} is an $n \times p$ matrix, then \mathbf{Ax} is a column vector. From the properties of matrix multiplication it can easily be shown that

$$\mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{Ax} + \mathbf{Ay}, \quad (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^p \quad (4.34)$$

and

$$\mathbf{A}(c\mathbf{x}) = c\mathbf{Ax}, \quad c \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^p. \quad (4.35)$$

Using this observation we can regard an $n \times p$ matrix as defining a *mapping* from \mathbb{R}^p to \mathbb{R}^n by the formula

$$\mathbf{A} : \mathbb{R}^p \rightarrow \mathbb{R}^n, \quad \mathbf{x} \rightarrow \mathbf{Ax} \quad (\rightarrow \text{stands for 'replaced by'}). \quad (4.36)$$

Properties (4.34) and (4.35) are usually called the *linearity* properties of \mathbf{A} and in this context \mathbf{A} is referred to as a *linear transformation* or *linear operator*. Conversely, it is easily shown that any linear operator $\mathbf{T} : \mathbb{R}^p \rightarrow \mathbb{R}^n$ can be represented by a matrix \mathbf{A} .

To see this, let $\mathbf{e}_i = (0, \dots, 1, \dots, 0)^T$ denote a column vector with one in the i -th position and zeros elsewhere. Then any column vector \mathbf{x} of dimension p can be written as

$$\mathbf{x} = \sum_{i=1}^p c_i \mathbf{e}_i. \quad (4.37)$$

Hence, if (4.37) satisfies (4.34)-(4.35) using these repeatedly gives

$$\mathbf{T}\mathbf{x} = \sum_{i=1}^p c_i \mathbf{T}\mathbf{e}_i. \quad (4.38)$$

Now, $\mathbf{T}\mathbf{e}_i \in \mathbb{R}^n$ so it is some column vector $\mathbf{a}_i = (a_{1i}, a_{2i}, \dots, a_{ni})^T$, $1 \leq i \leq p$. Defining \mathbf{A} as the matrix whose i -th column is \mathbf{a}_i it follows easily that

$$\mathbf{T}\mathbf{x} = \mathbf{A}\mathbf{x}. \quad (4.39)$$

Hence, we can think of a matrix and its corresponding linear transformation interchangeably. Often the geometric language is more convenient.

4.4.1 Matrix Inversion

If a is a nonzero real number, then a has a multiplicative inverse $1/a$ which has the property that $(1/a)a = a(1/a) = 1$. The existence of the multiplicative inverse of a nonzero real number then allows one to define an operation inverse to multiplication, division. In this section we take up this possibility for nonzero matrices.

As we shall see, the process of matrix inversion is crucial to the development of much of the estimation theory for linear regression and our development is largely motivated by that fact. Although it is possible to define inverses of rectangular matrices our treatment of linear statistical models only requires that we investigate this process for square matrices. Even here, the subject is somewhat more complicated than for real numbers, since an arbitrary nonzero matrix need not have an inverse. At this point we will give only the only basic definitions and facts concerning inverses. Further theorems and some computational aspects will be considered in Section 4.8.

For further reading on this aspect of matrix algebra and its relation to statistical calculations we recommend that the reader consult Refs. [45, 103].

Definition 4.1 Let \mathbf{A} be an $n \times n$ matrix. We say that the $n \times n$ matrix \mathbf{B} is an inverse for \mathbf{A} if and only if

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n. \quad (4.40)$$

Theorem 4.1 If \mathbf{A} has an inverse, then the inverse is unique and will be written as \mathbf{A}^{-1} . Thus,

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}_n. \quad (4.41)$$

Proof. Suppose that \mathbf{B} and \mathbf{C} are two inverses for \mathbf{A} , then since $\mathbf{AB} = \mathbf{I}_n$ and $\mathbf{CA} = \mathbf{I}_n$, $\mathbf{C}(\mathbf{AB}) = (\mathbf{CA})\mathbf{B} = \mathbf{I}_n\mathbf{B} = \mathbf{B} = \mathbf{CI}_n = \mathbf{C}$, so that $\mathbf{B} = \mathbf{C}$ and the inverse is unique. ■

For practical purposes it is important to know which, if any, matrices have an inverse. Unfortunately, this is not always easily ascertained and may, in practical situations, require a great deal of computation. We first note that although the zero matrix cannot have an inverse, not every nonzero matrix has one either. To see this, let

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

and suppose that

$$\mathbf{B} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is the inverse of \mathbf{A} . Then $\mathbf{AB} = \mathbf{I}_2$, so that

$$\mathbf{AB} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} = \mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

But this is impossible since $1 \neq 0$.

For a 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad (4.42)$$

one can easily show that \mathbf{A}^{-1} exists if and only if the *determinant* of \mathbf{A} , $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21} \neq 0$. In fact, the inverse of \mathbf{A} is given by

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \quad (4.43)$$

The Equation (4.43) is a particular case of *Cramer's rule* and it is well known that the determinant criterion and formula may be extended to $n \times n$ matrices. Since we will have little further use for such formulas the interested reader may consult Refs. [112, 103] for more information.

For theoretical purposes a more convenient criterion for determining the invertibility of \mathbf{A} is the following.

Theorem 4.2 *Let \mathbf{A} be an $n \times n$ matrix. Then \mathbf{A} has an inverse if and only if the only n -vector \mathbf{x} satisfying $\mathbf{Ax} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.*

A proof of Theorem 4.2 may be found in [112, 103]. It's usefulness stems from the fact that the equation $\mathbf{Ax} = \mathbf{0}$, when written out in full, is a system of n linear equations in n unknowns and standard algorithms, such as *Gaussian elimination*, may be used to solve them. If the only solution to that system is $\mathbf{0}$ we may conclude that \mathbf{A} has an inverse.

Example 4.1 Using Theorem 4.2 show that

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}$$

has an inverse.

Let $\mathbf{x} = (x, y, z)^T$, then we need only show that $\mathbf{Ax} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$. But,

$$\mathbf{Ax} = \begin{bmatrix} x + y + z \\ x + 2z \\ y + z \end{bmatrix}$$

so that $\mathbf{Ax} = \mathbf{0}$ gives

$$\begin{cases} x + y + z = 0 \\ x + 2z = 0 \\ y + z = 0 \end{cases}.$$

Now subtracting the second equation from the first gives $y - z = 0$ and using this and $y + z = 0$ gives $y = z = 0$. Substituting this into $x + y + z = 0$ we get $x = 0$, showing that $\mathbf{x} = (0, 0, 0)^T = \mathbf{0}$. Since $\mathbf{0}$ is the only solution to $\mathbf{Ax} = \mathbf{0}$, we conclude from Theorem 4.2 that \mathbf{A} has an inverse. (Notice that we did not need to compute \mathbf{A}^{-1} .)

Another convenient way of regarding Theorem 4.2 is to observe that

$$\mathbf{Ax} = \sum_{i=1}^n x_i \mathbf{a}_i \quad (4.44)$$

where \mathbf{a}_i , $1 \leq i \leq n$, is the i -th column of \mathbf{A} . Hence, the condition that $\mathbf{Ax} = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$ is the same as $\sum_{i=1}^n x_i \mathbf{a}_i = \mathbf{0} \Rightarrow x_i = 0, 1 \leq i \leq n$. And this says the vectors $\mathbf{a}_i, 1 \leq i \leq n$, are *linearly independent*. That is, the only *linear combination* of the column vectors of \mathbf{A} that can add up to zero must have all the coefficients $x_i, 1 \leq i \leq n$, equal to zero. Thus, if \mathbf{A} is not invertible, there must exist at least one nonzero vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ such that

$$\sum_{i=1}^n x_i \mathbf{a}_i = \mathbf{0}. \quad (4.45)$$

In this case we say that the column vectors are *linearly dependent*.

More generally, if \mathbf{A} is an $n \times p$ matrix we define the *rank* of \mathbf{A} to be the number of linearly independent columns of \mathbf{A} . So, if \mathbf{A} is an $n \times n$ matrix, then it is invertible if and only if its rank is n .

An important property of the rank is that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$ so that the rank of \mathbf{A} is also the number of linearly independent rows. Hence, for an $n \times p$ matrix $\text{rank}(\mathbf{A}) \leq \min(n, p)$. A matrix having the property that $\text{rank}(\mathbf{A}) = \min(n, p)$ is said to have *full rank*. Obviously, an invertible $n \times n$ matrix has rank n . These theoretical properties, although perhaps a little abstract for now, will find considerable use in later chapters.

Theorem 4.3 (Some further properties of \mathbf{A}^{-1})

(i) If \mathbf{A} and \mathbf{B} are invertible matrices, then \mathbf{AB} is invertible and

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}. \quad (4.46)$$

(ii) If \mathbf{A} is invertible, then $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$.

(iii) If $\mathbf{A} = \text{diag}(a_i), 1 \leq i \leq n$, and $a_i \neq 0, 1 \leq i \leq n$, then $\mathbf{A}^{-1} = \text{diag}(1/a_i), 1 \leq i \leq n$. In particular, $\mathbf{I}_n^{-1} = \mathbf{I}_n$.

(iv) \mathbf{A} is invertible if and only if $\det(\mathbf{A}) \neq 0$.

Proof. (i) It suffices to show that $(\mathbf{B}^{-1}\mathbf{A}^{-1})\mathbf{AB} = \mathbf{AB}(\mathbf{B}^{-1}\mathbf{A}^{-1}) = \mathbf{I}_n$. Now using associativity,

$$(\mathbf{B}^{-1}\mathbf{A}^{-1})\mathbf{AB} = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\mathbf{B} = \mathbf{B}^{-1}(\mathbf{I}_n)\mathbf{B} = \mathbf{BB}^{-1} = \mathbf{I}_n. \quad (4.47)$$

Similarly, $\mathbf{AB}(\mathbf{B}^{-1}\mathbf{A}^{-1}) = \mathbf{I}_n$.

(ii) Since $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$, taking the transpose of this gives $(\mathbf{A}^{-1}\mathbf{A})^T = (\mathbf{A}^T)(\mathbf{A}^{-1})^T = \mathbf{I}_n$. Similarly, $(\mathbf{A}^{-1})^T\mathbf{A}^T = \mathbf{I}_n$, which shows that $(\mathbf{A}^{-1})^T$ is the inverse of $(\mathbf{A}^T)^{-1}$.

(iii) We leave the proof of (iii) as an exercise and (iv) is a standard theorem of linear algebra which may be found in [112, 103]. ■

4.4.2 The Inverse of Partitioned Matrices

It is often convenient to regard a matrix \mathbf{A} in terms of various *submatrices*. For example, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad (4.48)$$

then we can consider it as composed of submatrices, delineated by bars, as follows;

$$\mathbf{A} = \left[\begin{array}{cc|cc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{array} \right] \quad (4.49)$$

and this can be further abbreviated as

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{B} & \mathbf{C} \\ \hline \mathbf{D} & \mathbf{E} \end{array} \right] \quad (4.50)$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are the submatrices indicated in (4.49). When \mathbf{A} is written in the form (4.50) we shall say it is in *partitioned* or *block form*.

If \mathbf{A} and \mathbf{B} are conformable matrices and each is partitioned as

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array} \right], \quad \mathbf{B} = \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array} \right] \quad (4.51)$$

then it can be shown that \mathbf{AB} can be written in partitioned form as

$$\mathbf{AB} = \left[\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array} \right] \quad (4.52)$$

provided that the indicated products are well defined. Note that this product is formed by the same rule as if the elements were scalars, keeping in mind that the correct order must be used.

Using (4.52) we can develop a number of formulas for the inversion of matrices in partitioned form.

Theorem 4.4 *If*

$$\mathbf{W} = \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array} \right] \quad (4.53)$$

and

$$\mathbf{P} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}, \quad \mathbf{Q} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \quad (4.54)$$

then

$$\mathbf{W}^{-1} = \left[\begin{array}{c|c} \mathbf{P}^{-1} & \mathbf{A}\mathbf{B}\mathbf{Q}^{-1} \\ \hline \mathbf{D}^{-1}\mathbf{C}\mathbf{P}^{-1} & \mathbf{Q}^{-1} \end{array} \right] \quad (4.55)$$

provided the requisite inverses exist.

Proof. Assuming \mathbf{W}^{-1} exists we write it in block form as

$$\mathbf{W}^{-1} = \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D}_1 \end{array} \right] \quad (4.56)$$

and then

$$\mathbf{W}^{-1}\mathbf{W} = \mathbf{I}_n = \left[\begin{array}{c|c} \mathbf{I}_p & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_q \end{array} \right] \quad (4.57)$$

where $p + q = n$. Multiplying (4.53) out gives

$$\mathbf{W}^{-1}\mathbf{W} = \left[\begin{array}{c|c} \mathbf{A}_1\mathbf{A} + \mathbf{B}_1\mathbf{C} & \mathbf{A}_1\mathbf{B} + \mathbf{B}_1\mathbf{D} \\ \hline \mathbf{C}_1\mathbf{A} + \mathbf{D}_1\mathbf{C} & \mathbf{C}_1\mathbf{B} + \mathbf{D}_1\mathbf{D} \end{array} \right]. \quad (4.58)$$

Equating matrices in (4.58) gives

$$\begin{cases} \mathbf{A}_1\mathbf{A} + \mathbf{B}_1\mathbf{C} = \mathbf{I}_p, & \mathbf{A}_1\mathbf{B} + \mathbf{B}_1\mathbf{D} = \mathbf{0}, \\ \mathbf{C}_1\mathbf{A} + \mathbf{D}_1\mathbf{C} = \mathbf{0}, & \mathbf{C}_1\mathbf{B} + \mathbf{D}_1\mathbf{D} = \mathbf{I}_q. \end{cases} \quad (4.59)$$

From (4.59)

$$\mathbf{A}_1 = (\mathbf{I}_p - \mathbf{B}_1\mathbf{C}) \mathbf{A}^{-1} \quad (4.60)$$

and using this in the second equation of (4.59) gives

$$(\mathbf{I} - \mathbf{B}_1\mathbf{C}) \mathbf{A}^{-1}\mathbf{B} + \mathbf{B}_1\mathbf{D} = \mathbf{0}. \quad (4.61)$$

Thus,

$$\mathbf{B}_1\mathbf{D} - \mathbf{B}_1\mathbf{C}\mathbf{A}^{-1}\mathbf{B} + \mathbf{A}^{-1}\mathbf{B} = \mathbf{B}_1(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) + \mathbf{A}^{-1}\mathbf{B} = \mathbf{0}. \quad (4.62)$$

Solving for \mathbf{B}_1 gives

$$\mathbf{B}_1 = -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{Q}^{-1}. \quad (4.63)$$

Using (4.59) again gives

$$\mathbf{B}_1\mathbf{D} = -\mathbf{A}_1\mathbf{B}. \quad (4.64)$$

So

$$\mathbf{B}_1 = -\mathbf{A}_1\mathbf{B}\mathbf{D}^{-1}. \quad (4.65)$$

Thus,

$$\mathbf{A}_1\mathbf{A} - \mathbf{A}_1\mathbf{B}\mathbf{D}^{-1}\mathbf{C} = \mathbf{I}_p \quad (4.66)$$

so that

$$\mathbf{A}_1(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) = \mathbf{I}_p \quad (4.67)$$

and solving for \mathbf{A}_1 gives

$$\mathbf{A}_1 = \mathbf{P}^{-1}. \quad (4.68)$$

Similarly we get \mathbf{C}_1 and \mathbf{D}_1 . We leave the details as an exercise. ■

4.4.3 The Sherman-Morrison-Woodbury Formula

In regression analysis one often needs to analyze the effect of adding or deleting a new observation. Doing this requires that we be able to calculate the inverse of a matrix perturbed by the addition of a rank one matrix. A famous formula given by Sherman, Morrison and Woodbury [8] (See also Rao (1973), p. 33) enables one to do this in terms of \mathbf{A}^{-1} . A proof of this is given next.

Theorem 4.5 (Sherman-Morrison-Woodbury Theorem) *Let \mathbf{A} be an $n \times n$ invertible matrix and let \mathbf{z} be an $n \times 1$ column vector. If $\mathbf{z}^T \mathbf{A}^{-1} \mathbf{z} \neq 1$ then the matrix $\mathbf{B} = \mathbf{A} - \mathbf{z}\mathbf{z}^T$ has an inverse and*

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1} \mathbf{z} \mathbf{z}^T \mathbf{A}^{-1}}{1 - \mathbf{z}^T \mathbf{A}^{-1} \mathbf{z}}. \quad (4.69)$$

Before proving the theorem we establish another result in matrix multiplication which will be needed in the course of the proof.

Let \mathbf{u}, \mathbf{v} and \mathbf{y} be $n \times 1$ column vectors, then

$$(\mathbf{u}^T \mathbf{y}) \mathbf{v} = (\mathbf{v} \mathbf{u}^T) \mathbf{y}. \quad (4.70)$$

To see this, we write $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)^T$. Then,

$$\mathbf{u}^T \mathbf{y} = (u_1, u_2, \dots, u_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^n u_i y_i \quad (4.71)$$

so that

$$(\mathbf{u}^T \mathbf{y}) \mathbf{v} = \left[\left(\sum_{i=1}^n u_i y_i \right) v_1, \left(\sum_{i=1}^n u_i y_i \right) v_2, \dots, \left(\sum_{i=1}^n u_i y_i \right) v_n \right]^T. \quad (4.72)$$

Also,

$$\mathbf{v} \mathbf{u}^T = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \cdot (u_1, u_2, \dots, u_n) = \begin{bmatrix} v_1 u_1 & v_1 u_2 & \cdots & v_1 u_n \\ v_2 u_1 & v_2 u_2 & \cdots & v_2 u_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n u_1 & v_n u_2 & \cdots & v_n u_n \end{bmatrix} \quad (4.73)$$

so that

$$\begin{aligned} (\mathbf{v} \mathbf{u}^T) \mathbf{y} &= \begin{bmatrix} v_1 u_1 & v_1 u_2 & \cdots & v_1 u_n \\ v_2 u_1 & v_2 u_2 & \cdots & v_2 u_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n u_1 & v_n u_2 & \cdots & v_n u_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} v_1 (\sum_{i=1}^n u_i y_i) \\ v_2 (\sum_{i=1}^n u_i y_i) \\ \vdots \\ v_n (\sum_{i=1}^n u_i y_i) \end{bmatrix} \\ &= (\mathbf{u}^T \mathbf{y}) \mathbf{v}. \end{aligned} \quad (4.74)$$

Proof of Theorem 4.5. We begin by observing that if the set of equations $\mathbf{C}\mathbf{x} = \mathbf{y}$ has a solution $\mathbf{x} = \mathbf{D}\mathbf{y}$ for all n -vectors \mathbf{y} , then $\mathbf{D} = \mathbf{C}^{-1}$. (This follows from the uniqueness of the matrix inverse.) Hence, we consider solving

$$\mathbf{B}\mathbf{x} = \mathbf{y}, \mathbf{y} \in \mathbb{R}^n. \quad (4.75)$$

Using the definition of \mathbf{B} , and the above observation

$$\mathbf{B}\mathbf{x} = \mathbf{A}\mathbf{x} - \mathbf{z}\mathbf{z}^T\mathbf{x} = \mathbf{A}\mathbf{x} - c\mathbf{z} \quad (4.76)$$

where $c = \mathbf{z}^T\mathbf{x}$ (which is a scalar). Thus,

$$\mathbf{A}\mathbf{x} - c\mathbf{z} = \mathbf{y} \quad (4.77)$$

so that

$$\mathbf{x} = c\mathbf{A}^{-1}\mathbf{z} + \mathbf{A}^{-1}\mathbf{y}. \quad (4.78)$$

Multiplying both sides of (4.78) by \mathbf{z}^T gives

$$\mathbf{z}^T\mathbf{x} = c = c\mathbf{z}^T\mathbf{A}^{-1}\mathbf{z} + \mathbf{z}^T\mathbf{A}^{-1}\mathbf{y} \quad (4.79)$$

and solving for c we get

$$c = \frac{\mathbf{z}^T\mathbf{A}^{-1}\mathbf{y}}{1 - \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}}. \quad (4.80)$$

Substituting this value of c into (4.78) gives

$$\mathbf{x} = \frac{(\mathbf{z}^T\mathbf{A}^{-1}\mathbf{y})\mathbf{A}^{-1}\mathbf{z}}{1 - \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}} + \mathbf{A}^{-1}\mathbf{y}. \quad (4.81)$$

Letting $\mathbf{u}^T = \mathbf{z}^T\mathbf{A}^{-1}$ and $\mathbf{v} = \mathbf{A}^{-1}\mathbf{z}$ in (4.70), then

$$(\mathbf{z}^T\mathbf{A}^{-1}\mathbf{y})\mathbf{A}^{-1}\mathbf{z} = (\mathbf{A}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{A}^{-1})\mathbf{y}. \quad (4.82)$$

Thus,

$$\mathbf{x} = \mathbf{C}\mathbf{y} \quad (4.83)$$

where

$$\mathbf{C} = \frac{\mathbf{A}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{A}^{-1}}{1 - \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}} + \mathbf{A}^{-1}. \quad (4.84)$$

From our remark at the beginning of the proof $\mathbf{C} = \mathbf{B}^{-1}$. ■

4.5 The Geometry of Vectors

It will be convenient to interpret some aspects of matrix and vector algebra in geometric terms. Just as a 2-vector $(x, y)^T$ may be interpreted as a point in the plane or as a line segment (geometric vector) joining the origin $(0, 0)^T$ to the point $(x, y)^T$ with a direction pointing from $(0, 0)^T$ to $(x, y)^T$, an n -vector $(x_1, x_2, \dots, x_n)^T$ may be interpreted as a point in Euclidean n -space \mathbb{R}^n or as a geometric vector pointing from the origin $(0, 0, \dots, 0)^T$ to $(x_1, x_2, \dots, x_n)^T$. With this geometric interpretation of vectors we can introduce the

important notations of length and angle. Basic to doing this is the notation of the *inner product* of two vectors.

Definition 4.2 Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ be two n -vectors. The *inner product* $\langle \mathbf{x}, \mathbf{y} \rangle$ (sometimes also called the *dot product*) of \mathbf{x} and \mathbf{y} is the real number defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle \equiv \sum_{i=1}^n x_i y_i. \quad (4.85)$$

Note: If \mathbf{y} is regarded as a row, rather than as a column vector, then $\langle \mathbf{x}, \mathbf{y} \rangle$ may be written in matrix multiplication notation as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T \mathbf{x} = (y_1, y_2, \dots, y_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad (4.86)$$

which is preferred by many authors.

The inner product has a number of simple properties which we quote below. Because of their simplicity, we leave most of the proofs of these to the reader.

Theorem 4.6 (Properties of the inner product)

- (i) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (*symmetry*);
- (ii) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$;
- (iii) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
- (iv) If c is a real number, then

$$\langle c\mathbf{x}, \mathbf{y} \rangle = c \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, c\mathbf{y} \rangle. \quad (4.87)$$

- (v) If \mathbf{A} is an $m \times n$ matrix, then

$$\langle \mathbf{x}, \mathbf{A}\mathbf{y} \rangle = \langle \mathbf{A}^T \mathbf{x}, \mathbf{y} \rangle. \quad (4.88)$$

- (vi) (*The Cauchy-Schwarz inequality*)

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}. \quad (4.89)$$

Proof. (i)-(v) require only straightforward algebraic manipulations and are left as exercises.

For (vi) we consider the quantity $Q(\lambda) = \langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y} \rangle$. Expanding $Q(\lambda)$ using (i)-(iv) of the theorem gives

$$Q(\lambda) = \langle \mathbf{x}, \mathbf{x} \rangle + 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + \lambda^2 \langle \mathbf{y}, \mathbf{y} \rangle \quad (4.90)$$

and from (4.90) we see that $Q(\lambda)$ is a quadratic function of λ and $Q(\lambda) \geq 0$. We consider two cases:

- (a) if $\mathbf{y} = 0$, then $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ and $\langle \mathbf{y}, \mathbf{y} \rangle = 0$ so that (4.90) is true
- (b) if $\mathbf{y} \neq 0$, then $Q(\lambda) = 0$ is a true quadratic and from elementary algebra the only way that $Q(\lambda) \geq 0$ is that $Q(\lambda)$ either has no real roots or two real equal roots. In these cases the discriminant

$$4 \langle \mathbf{x}, \mathbf{y} \rangle^2 - 4 \langle \mathbf{y}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{x} \rangle \leq 0. \quad (4.91)$$

Transposing and taking square roots gives

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}.$$

■

4.5.1 Length and Angle

From elementary geometry (Pythagorean theorem) it is well known that the length of the vector in two dimensions $\mathbf{x} = (x_1, x_2)^T$ is given by

$$\sqrt{x_1^2 + x_2^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \quad (4.92)$$

Generalizing this for a vector with n components we define the *length* of $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ by

$$\|\mathbf{x}\| \equiv \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}. \quad (4.93)$$

Using Theorem 4.6 $\|\mathbf{x}\|$ can be shown to have the expected properties of length. Among these are:

- (i) $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$;
- (ii) $\|c\mathbf{x}\| = |c| \|\mathbf{x}\|$;
- (iii) (*Triangle inequality*) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

Proof. Properties (i) and (ii) being elementary, we establish (iii). For this we consider

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + 2 \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 \langle \mathbf{x}, \mathbf{y} \rangle. \end{aligned} \quad (4.94)$$

By the Cauchy-Schwarz inequality $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$ so that

$$\|\mathbf{x} + \mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 \|\mathbf{x}\| \|\mathbf{y}\| = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \quad (4.95)$$

Taking square roots in (4.95) gives

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad (4.96)$$

as required. ■

Along with length, the inner product allows us to introduce the concept of the angle between two vectors. For this, assume that $\mathbf{x} \neq 0$, $\mathbf{y} \neq 0$ and consider the ratio

$$r = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (4.97)$$

From the Cauchy-Schwarz inequality $-1 \leq r \leq 1$, so from trigonometry there is an angle $0 \leq \theta \leq \pi$ such that

$$\cos \theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (4.98)$$

Since the cosine is single-valued on $[0, \pi]$, we have

$$\theta = \cos^{-1} \left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} \right). \quad (4.99)$$

We call θ the *angle* between the vectors \mathbf{x} and \mathbf{y} . If either \mathbf{x} or \mathbf{y} is zero, we leave the notion of angle undefined. If $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, then $\theta = \cos^{-1}(0) = \pi/2$ and we then say that \mathbf{x} and \mathbf{y} are *orthogonal vectors*. Orthogonality generalizes the notion of perpendicularity of vectors in the plane and three-space.

As we shall see, orthogonality is a fundamental concept in much of regression analysis and much of its importance stems from the following geometric fact. Suppose that we have a line \mathbf{l} in the plane that passes through the origin. Such a line can be given in geometric terms as the set of vectors $\{t\mathbf{v}\}$, $t \in \mathbb{R}$, where \mathbf{v} is a vector of unit length lying in \mathbf{l} . Now consider a point (p, q) which does not lie on \mathbf{l} and consider the point on \mathbf{l} which is closest to \mathbf{l} . From Figure 4.2 it is clear that this point, call it $t_1\mathbf{v}$, has the property that the vector $t_1\mathbf{v}$ must be perpendicular to the line passing through the point $\mathbf{y} = (p, q)$ and $(t_1\mathbf{v})$. Thus the vectors $\mathbf{y} - t_1\mathbf{v}$ and $t_1\mathbf{v}$ must be orthogonal

$$\langle \mathbf{y} - t_1\mathbf{v}, t_1\mathbf{v} \rangle = 0 \quad (4.100)$$

and solving for t_1 using $\langle \mathbf{v}, \mathbf{v} \rangle = 1$ gives

$$t_1 = \langle \mathbf{y}, \mathbf{v} \rangle. \quad (4.101)$$

The vector

$$\mathbf{y}_p = \langle \mathbf{y}, \mathbf{v} \rangle \mathbf{v} \quad (4.102)$$

is called the *orthogonal projection* of \mathbf{y} onto the line \mathbf{l} . Suitably generalized this construction is the geometrical basis for the method of least squares which we shall discuss in greater detail in Chapter 5.

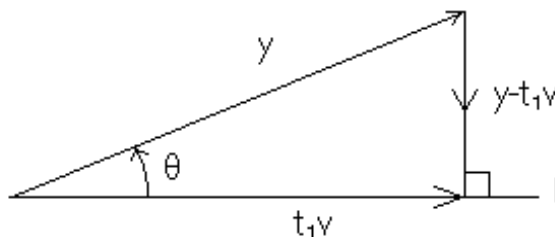


Figure 4.2: The Orthogonal Projection of \mathbf{y} on \mathbf{l}

In that case we will consider the set of n -vectors $\mathcal{P} = \left\{ \sum_{j=1}^p t_j \mathbf{v}_j \right\}$ called a *hyperplane* where $\{\mathbf{v}_j\}_{j=1}^p$ are linearly independent and $p \leq n$. Again we suppose that $\mathbf{y} \in \mathbb{R}^n$ and we want to find the point in the hyperplane \mathcal{P} closest to \mathbf{y} . Generalizing the geometry of the two dimensional case, this will occur when the vector $\mathbf{y} - \sum_{j=1}^p t_j \mathbf{v}_j$ is orthogonal to \mathcal{P} , and this is equivalent to having $\mathbf{y} - \sum_{j=1}^p t_j \mathbf{v}_j$ perpendicular to each $\mathbf{v}_j, j = 1, 2, \dots, p$. Thus, the coordinates $\{t_i\}$ are found by solving the linear equations

$$\sum_{j=1}^p \langle \mathbf{v}_i, \mathbf{v}_j \rangle t_j = \langle \mathbf{y}, \mathbf{v}_i \rangle, \quad 1 \leq i \leq p. \quad (4.103)$$

As we shall see, these are the generalizations of the least squares equations used for estimating the parameters in the simple regression model.

4.5.2 Subspaces and Bases

As with vectors, it is often convenient to discuss matrices in geometric terms. As we have already observed, we can view a matrix as composed of columns $\mathbf{a}_i, 1 \leq i \leq p$. If we consider the set of linear combinations of $\mathbf{a}_i, 1 \leq i \leq p$,

$$\mathbb{S} = \left\{ \sum_{j=1}^p t_j \mathbf{a}_j \mid t_j \in \mathbb{R}, 1 \leq j \leq p \right\} \quad (4.104)$$

then these determine a hyperplane in \mathbb{R}^n called the *column space spanned* by $\mathbf{a}_i, 1 \leq i \leq p$. More generally, if we have any sequence of vectors $\mathbf{x}_i, 1 \leq i \leq p$ in \mathbb{R}^n the set of all linear combinations of $\mathbf{x}_i, 1 \leq i \leq p$, is called a *subspace* of \mathbb{R}^n . One can easily show that if \mathbb{S} is a subspace, then it is closed under the operations of scalar multiplication and addition i.e., if $\mathbf{x} \in \mathbb{S}$ and $c \in \mathbb{R}$ then $c\mathbf{x} \in \mathbb{S}$ and if $\mathbf{x} \in \mathbb{S}$ and $\mathbf{y} \in \mathbb{S}$, then $\mathbf{x} + \mathbf{y} \in \mathbb{S}$. Conversely, it can be shown that if \mathbb{S} is any subset of \mathbb{R}^n closed under scalar multiplication and addition, then there is a set of vectors $\{\mathbf{x}_i\}_{i=1}^p$ such that \mathbb{S} is *spanned* by the \mathbf{x}_i 's i.e., it can be written in the form (4.104). Since the set $\{\mathbf{x}_i\}_{i=1}^p$ can contain redundant vectors, we consider the smallest subset of the \mathbf{x}_i 's which spans \mathbb{S} . It can be shown that such a set of vectors is linearly independent. Such a set of linearly independent spanning vectors is called a *basis* for \mathbb{S} . An important theorem in linear algebra states that all bases for a subspace \mathbb{S} contain the same number of vectors. This number is referred to as the *dimension* of \mathbb{S} , denoted by $\dim(\mathbb{S})$. For a matrix \mathbf{A} the dimension of the column space is the rank of \mathbf{A} .

In general, if \mathbb{S}_1 and \mathbb{S}_2 are two subspaces of \mathbb{R}^n and $\mathbb{S}_1 \subseteq \mathbb{S}_2$, then $\dim(\mathbb{S}_1) \leq \dim(\mathbb{S}_2)$. In particular, $\dim(\mathbb{R}^n) = n$ since \mathbb{R}^n is spanned by the *canonical basis*

$$\mathbf{e}_i = \left(0, \dots, 0, \underset{i\text{-th}}{1}, 0, \dots, 0 \right)^T, \quad 1 \leq i \leq n. \quad (4.105)$$

The canonical basis $\{\mathbf{e}_i\}_{i=1}^n$ of \mathbb{R}^n has an important additional property, it is an *orthogonal basis*; i.e.,

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0, \quad i \neq j. \quad (4.106)$$

Orthogonal bases play an important role in regression analysis and it is important to know if a given subspace has an orthogonal basis. Fortunately, the answer is 'yes'.

Moreover, this property can be established constructively by a process called *Gram-Schmidt orthogonalization*, which is important both theoretically and computationally.

Theorem 4.7 *Let \mathbb{S} be a subspace of \mathbb{R}^n . If $\dim(\mathbb{S}) = p \leq n$, then \mathbb{S} has an orthogonal basis.*

Proof. A full proof of the theorem can be found in [112]. However, we merely outline the general construction, called, as noted above, Gram-Schmidt orthogonalization.

Since \mathbb{S} is a subspace it has a basis $\mathbf{v}_i, 1 \leq i \leq p$. If the \mathbf{v}_i 's are orthogonal, then we are done. If not, we construct an orthogonal basis $\mathbf{w}_i, 1 \leq i \leq p$, from the \mathbf{v}_i 's which spans \mathbb{S} . Since orthogonal vectors are linearly independent, $\mathbf{w}_i, 1 \leq i \leq p$, are a basis for \mathbb{S} .

The basic idea of the proof is to start with \mathbf{v}_1 , and normalize it by dividing by $\|\mathbf{v}_1\|$ giving $\mathbf{w}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|$, (notice that $\|\mathbf{w}_1\| = 1$). Then at each stage having obtained $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l$ we compute the orthogonal projection, $\text{Proj}_{\mathbf{v}_{l+1}}$ of \mathbf{v}_{l+1} onto the span of $\{\mathbf{w}_j\}_{j=1}^l$ and then form

$$\mathbf{w}'_{l+1} = \mathbf{v}_{l+1} - \text{Proj}_{\mathbf{v}_{l+1}}. \quad (4.107)$$

Finally,

$$\mathbf{w}_{l+1} = \frac{\mathbf{w}'_{l+1}}{\|\mathbf{w}'_{l+1}\|}. \quad (4.108)$$

We demonstrate this explicitly for vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. The projection of \mathbf{v}_2 on \mathbf{w}_1 is given by $\langle \mathbf{v}_2, \mathbf{w}_1 \rangle \mathbf{w}_1$, then

$$\mathbf{w}'_2 = \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{w}_1 \rangle \mathbf{w}_1 \quad (4.109)$$

so that

$$\begin{aligned} \langle \mathbf{w}'_2, \mathbf{w}_1 \rangle &= \langle \mathbf{w}_1, \mathbf{v}_2 \rangle - \langle \mathbf{v}_2, \mathbf{w}_1 \rangle \langle \mathbf{w}_1, \mathbf{w}_1 \rangle \\ &= \langle \mathbf{w}_1, \mathbf{v}_2 \rangle - \langle \mathbf{v}_2, \mathbf{w}_1 \rangle = 0. \end{aligned} \quad (4.110)$$

Hence, \mathbf{w}_1 and \mathbf{w}'_2 are orthogonal. Then, $\mathbf{w}_2 = \mathbf{w}'_2 / \|\mathbf{w}'_2\|$. Following (4.109) we define \mathbf{w}'_3 by

$$\mathbf{w}'_3 = \mathbf{v}_3 - \langle \mathbf{v}_3, \mathbf{w}_1 \rangle \mathbf{w}_1 - \langle \mathbf{v}_3, \mathbf{w}_2 \rangle \mathbf{w}_2. \quad (4.111)$$

From (4.110)

$$\begin{aligned} \langle \mathbf{w}'_3, \mathbf{w}_1 \rangle &= \langle \mathbf{v}_3, \mathbf{w}_1 \rangle - \langle \mathbf{v}_3, \mathbf{w}_1 \rangle \langle \mathbf{w}_1, \mathbf{w}_1 \rangle - \langle \mathbf{v}_3, \mathbf{w}_2 \rangle \langle \mathbf{w}_2, \mathbf{w}_1 \rangle \\ &= \langle \mathbf{v}_3, \mathbf{w}_1 \rangle - \langle \mathbf{v}_3, \mathbf{w}_1 \rangle = 0 \end{aligned} \quad (4.112)$$

since \mathbf{w}_1 and \mathbf{w}_2 are orthogonal. Similarly, $\langle \mathbf{w}'_3, \mathbf{w}_2 \rangle = 0$. Then, $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 = \mathbf{w}'_3 / \|\mathbf{w}'_3\|$ are an orthogonal basis for $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$.

In general, it follows from (4.112) that \mathbf{w}'_{l+1} in (4.108) is given explicitly by

$$\mathbf{w}'_{l+1} = \mathbf{v}_{l+1} - \sum_{j=1}^l \langle \mathbf{v}_{l+1}, \mathbf{w}_j \rangle \mathbf{w}_j. \quad (4.113)$$

We leave it as an exercise to show that $\mathbf{w}'_j, 1 \leq j \leq p$, are orthogonal. ■

4.6 Orthogonal Matrices

We note that the vectors \mathbf{w}_i , $1 \leq i \leq p$, in Theorem 4.7 are not only orthogonal, but have the additional property that $\|\mathbf{w}_i\| = 1$, $1 \leq i \leq p$. Any set of vectors \mathbf{u}_i , $1 \leq i \leq p$, in \mathbb{R}^n satisfying

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \quad (4.114)$$

is said to be *orthonormal*.

If $p = n$, then $\{\mathbf{u}_i\}_{i=1}^n$ is an *orthonormal basis* for \mathbb{R}^n . If \mathbf{u}_i , $1 \leq i \leq n$, is an orthonormal basis for \mathbb{R}^n then the matrix

$$\mathbf{U} = [\mathbf{u}_1 | \mathbf{u}_2 | \cdots | \mathbf{u}_n] \quad (4.115)$$

whose i -th column is \mathbf{u}_i is called an *orthogonal matrix*. Orthogonal matrices will play a significant role in this and later chapters.

Theorem 4.8 (Properties of orthogonal matrices)

- (i) A matrix \mathbf{U} is orthogonal if and only if $\mathbf{U}^{-1} = \mathbf{U}^T$.
- (ii) If \mathbf{U} is orthogonal, then \mathbf{U}^T is orthogonal.
- (iii) If \mathbf{U}_1 and \mathbf{U}_2 are orthogonal, then $\mathbf{U}_1 \mathbf{U}_2$ is orthogonal.
- (iv) If \mathbf{U} is an $n \times n$ orthogonal matrix, then for all $\mathbf{x} \in \mathbb{R}^n$

$$\langle \mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle. \quad (4.116)$$

(Thus, the length of $\mathbf{U}\mathbf{x}$ is the same as \mathbf{x} . Hence, orthogonal matrices represent rotations in \mathbb{R}^n .)

- (v) If \mathbf{U} is orthogonal, then

$$\det(\mathbf{U}) = \pm 1. \quad (4.117)$$

Proof. (i) Consider the product $\mathbf{U}\mathbf{U}^T$. Then, by properties of matrix multiplication it is easily shown that the ij -th element of $\mathbf{U}\mathbf{U}^T$ is $\langle \mathbf{u}_i, \mathbf{u}_j \rangle$ because the j -th column of \mathbf{U}^T is the j -th row of \mathbf{U} . However, by assumption (4.114) $\{\mathbf{u}_i\}_{i=1}^n$ are orthogonal so that $\mathbf{U}\mathbf{U}^T = \mathbf{I}_n$. Similarly, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$ which shows that $\mathbf{U}^T = \mathbf{U}^{-1}$.

On the other hand, if $\mathbf{U}^{-1} = \mathbf{U}^T$, then $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}_n$ and (4.114) holds. Thus, an orthogonal matrix may be characterized by the property that $\mathbf{U}^{-1} = \mathbf{U}^T$.

(ii) Since \mathbf{U} is orthogonal, $\mathbf{U}^T = \mathbf{U}^{-1}$ and taking transpose gives $(\mathbf{U}^T)^T = \mathbf{U} = (\mathbf{U}^{-1})^T$. But $\mathbf{U} = (\mathbf{U}^{-1})^{-1}$ so that $(\mathbf{U}^{-1})^{-1} = (\mathbf{U}^{-1})^T$ which shows that \mathbf{U}^{-1} is orthogonal.

(iii) It suffices to show that $(\mathbf{U}_1 \mathbf{U}_2)^T = (\mathbf{U}_1 \mathbf{U}_2)^{-1}$. But

$$(\mathbf{U}_1 \mathbf{U}_2)^T = \mathbf{U}_2^T \mathbf{U}_1^T = \mathbf{U}_2^{-1} \mathbf{U}_1^{-1} = (\mathbf{U}_1 \mathbf{U}_2)^{-1}$$

by (4.21) and (i). Thus, $\mathbf{U}_1 \mathbf{U}_2$ is orthogonal if \mathbf{U}_1 and \mathbf{U}_2 are orthogonal.

(iv) For this we observe from (4.88) that

$$\langle \mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{U}^T \mathbf{U}\mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle$$

since $\mathbf{U}^T \mathbf{U} = \mathbf{I}_n$. (We note in passing that the converse of this result is also true. That is, if $\langle \mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle$ for all $\mathbf{x} \in \mathbb{R}^n$, then \mathbf{U} is orthogonal.)

(v) By a well known theorem on determinants $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$, so that

$$\det(\mathbf{UU}^T) = \det(\mathbf{I}_n) = \det(\mathbf{U}) \det(\mathbf{U}^T) = \det^2(\mathbf{U}) \quad (4.118)$$

since $\det(\mathbf{A}) = \det(\mathbf{A}^T)$ for all $n \times n$ matrices. But $\det(\mathbf{I}_n) = 1$, so that $\det^2(\mathbf{U}) = 1$ and so $\det(\mathbf{U}) = \pm 1$. ■

4.6.1 Eigenvectors and Eigenvalues

Since an arbitrary matrix can be a quite complicated mathematical object, it is often quite helpful, for both theory and computation, to be able to decompose a given matrix into simpler components, somewhat like factoring an integer into prime numbers. In matrix algebra many such decompositions exist, depending on the matrix involved.

In regression analysis decompositions of symmetric matrices are crucial to much of the theory. For us, the most important decomposition is a result of the so called *diagonalization* or *spectral theorem* which enables one to write a symmetric matrix \mathbf{A} as a product of two orthogonal and one diagonal matrix. The columns of the orthogonal component are composed of an important set of vectors related to \mathbf{A} , its *eigenvectors*.

We now briefly take up this important topic.

Definition 4.3 Let \mathbf{A} be an $n \times n$ matrix. We say that a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ is an *eigenvector* of \mathbf{A} if there exists a real number $\lambda \in \mathbb{R}$ such that

$$\mathbf{Ax} = \lambda \mathbf{x}. \quad (4.119)$$

The number λ is called an *eigenvalue* of \mathbf{A} corresponding to \mathbf{x} .

From a theoretical point of view, one of the problems associated with Definition 4.3 is that not every matrix \mathbf{A} need possess an eigenvector. Geometrically, this may be seen if we think of an eigenvector as defining a line which gets mapped into itself under \mathbf{A} . If, for example, \mathbf{A} is a matrix corresponding to a rotation, then no line remains fixed and so \mathbf{A} can have no eigenvector.

For example, if

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

which represents a clockwise rotation of 90° then $\mathbf{Ax} = (y, -x)^T$ so that $\mathbf{A}^2\mathbf{x} = (-x, -y)^T = -\mathbf{x}$. Thus, if \mathbf{x} is an eigenvector with eigenvalue λ , then $\mathbf{Ax} = \lambda\mathbf{x}$ so that $\mathbf{A}^2\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x} = -\mathbf{x}$ which gives $(\lambda^2 + 1)\mathbf{x} = 0$. If $\mathbf{x} \neq 0$, then $\lambda^2 + 1 = 0$ and there is no real value of λ satisfying this equation.

In order to alleviate some of the difficulties associated with this problem it is more convenient to permit both \mathbf{x} and λ to be complex. In this case, as is shown next, every matrix has at least one eigenvector-eigenvalue pair.

Theorem 4.9 Let \mathbf{A} be an $n \times n$ real matrix. Then, there exists at least one complex number λ and complex vector \mathbf{x} such that $\mathbf{Ax} = \lambda\mathbf{x}$.

Proof. For \mathbf{x} to be an eigenvector of \mathbf{A} we must have $(\mathbf{A} - \lambda \mathbf{I}_n) \mathbf{x} = 0$. Thus from Theorem 4.5 the matrix $\mathbf{A} - \lambda \mathbf{I}_n$ cannot have an inverse. From the theory of determinants, this can happen if and only if $c(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}_n) = 0$. But as is shown in [112], $\det(\mathbf{A} - \lambda \mathbf{I}_n)$ is a polynomial of *exact* degree n and from the *fundamental theorem of algebra* [112] this polynomial (called the *characteristic polynomial* of \mathbf{A}) must have at least one complex root λ . This root and a vector \mathbf{x} such that $(\mathbf{A} - \lambda \mathbf{I}_n) \mathbf{x} = 0$ is an eigenvector-eigenvalue pair for \mathbf{A} . ■

Since Theorem 4.9 shows that \mathbf{A} has at least one eigenvalue and these eigenvalues are the roots of $c(\lambda) = 0$, then an $n \times n$ matrix can have at most n distinct eigenvalues. However, it is possible for \mathbf{A} to have fewer than n distinct eigenvalues. Moreover, since any scalar multiple of an eigenvector is also an eigenvector, the question of classifying the nature of the eigenvectors is even more complex and will not be dealt with in generality here. However, if \mathbf{A} is symmetric, then things simplify considerably, and fortunately for regression analysis, this is the case of most importance. We take this up next.

4.6.2 The Spectral Theorem for Symmetric Matrices

Theorem 4.10 *If \mathbf{A} is an $n \times n$ symmetric matrix, then*

- (i) *all the eigenvalues of \mathbf{A} are real;*
- (ii) *eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Before proving Theorem 4.10 we need to extend the notion of an inner product to complex n -vectors. If $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ is a vector with each component a complex number, then $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)^T$ is called the *complex conjugate* of \mathbf{x} , where \bar{x}_i is the complex conjugate of x_i . The inner product of two complex vectors \mathbf{x}, \mathbf{y} , is then given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n x_i \bar{y}_i \quad (4.120)$$

which reduces to (4.85) if both \mathbf{x} and \mathbf{y} are real. The properties of the complex inner product are similar to those of Theorem 4.2 and are listed below.

- (i) $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$;
- (ii) $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$;
- (iii) $\langle \mathbf{x}, c\mathbf{y} \rangle = \bar{c} \langle \mathbf{x}, \mathbf{y} \rangle$, where c is a complex number and $\langle c\mathbf{x}, \mathbf{y} \rangle = c \langle \mathbf{x}, \mathbf{y} \rangle$;
- (iv) $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$, where $\|\mathbf{x}\| = (\sum_{i=1}^n x_i \bar{x}_i)^{1/2}$ and $\|\mathbf{y}\| = (\sum_{i=1}^n y_i \bar{y}_i)^{1/2}$.

Proof of Theorem 4.10. (i) Since λ is an eigenvalue (possibly complex) of \mathbf{A} , then $\mathbf{Ax} = \lambda \mathbf{x}$. Thus, $\langle \mathbf{x}, \mathbf{Ax} \rangle = \langle \mathbf{x}, \lambda \mathbf{x} \rangle$. From the property (iii) of the complex inner product $\langle \mathbf{x}, \lambda \mathbf{x} \rangle = \bar{\lambda} \langle \mathbf{x}, \mathbf{x} \rangle$ and using the fact that \mathbf{A} is real,

$$\langle \mathbf{x}, \mathbf{Ax} \rangle = \langle \mathbf{A}^T \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{Ax}, \mathbf{x} \rangle = \langle \lambda \mathbf{x}, \mathbf{x} \rangle = \lambda \langle \mathbf{x}, \mathbf{x} \rangle. \quad (4.121)$$

Thus, $\lambda \langle \mathbf{x}, \mathbf{x} \rangle = \bar{\lambda} \langle \mathbf{x}, \mathbf{x} \rangle$ and since $\mathbf{x} \neq 0$, $\langle \mathbf{x}, \mathbf{x} \rangle \neq 0$, which gives $\lambda = \bar{\lambda}$. But a complex number equals its complex conjugate if and only if it is real, so λ is real. (We note from this, that since $\mathbf{Ax} = \lambda \mathbf{x}$, that \mathbf{x} may be chosen to be real as well.)

(ii) Suppose that $\mathbf{Ax}_1 = \lambda_1 \mathbf{x}_1$ and $\mathbf{Ax}_2 = \lambda_2 \mathbf{x}_2$ where $\lambda_1 \neq \lambda_2$. Then, $\langle \mathbf{Ax}_1, \mathbf{x}_2 \rangle = \langle \lambda_1 \mathbf{x}_1, \mathbf{x}_2 \rangle$ and $\langle \mathbf{Ax}_2, \mathbf{x}_1 \rangle = \langle \lambda_2 \mathbf{x}_2, \mathbf{x}_1 \rangle$. Now, since \mathbf{x}_1 and \mathbf{x}_2 are real,

$$\langle \mathbf{Ax}_1, \mathbf{x}_2 \rangle = \langle \mathbf{x}_1, \mathbf{A}_2^T \mathbf{x} \rangle = \langle \mathbf{x}_1, \mathbf{Ax}_2 \rangle = \langle \mathbf{Ax}_2, \mathbf{x}_1 \rangle. \quad (4.122)$$

Thus, subtracting gives

$$\langle \lambda_1 \mathbf{x}_1, \mathbf{x}_2 \rangle - \langle \lambda_2 \mathbf{x}_2, \mathbf{x}_1 \rangle = \lambda_1 \langle \mathbf{x}_1, \mathbf{x}_2 \rangle - \lambda_2 \langle \mathbf{x}_2, \mathbf{x}_1 \rangle = (\lambda_1 - \lambda_2) \langle \mathbf{x}_1, \mathbf{x}_2 \rangle = 0.$$

Since $\lambda_1 \neq \lambda_2$, $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = 0$ showing that \mathbf{x}_1 and \mathbf{x}_2 are orthogonal. ■

As we pointed out above, an $n \times n$ matrix \mathbf{A} can have anywhere from one to n linearly independent eigenvectors. We will now examine what kinds of matrices can have n linearly independent eigenvectors.

Suppose this is true for \mathbf{A} , then there exist n linearly independent eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ which are not necessarily distinct. Thus, $\mathbf{Ax}_i = \lambda_i \mathbf{x}_i, 1 \leq i \leq n$. Let \mathbf{U} denote the matrix whose columns are $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ arranged in that order. Thus, $\mathbf{U} = [\mathbf{x}_1 | \mathbf{x}_2 | \dots | \mathbf{x}_n]$ and $\mathbf{AU} = [\mathbf{Ax}_1 | \mathbf{Ax}_2 | \dots | \mathbf{Ax}_n] = [\lambda_1 \mathbf{x}_1 | \lambda_2 \mathbf{x}_2 | \dots | \lambda_n \mathbf{x}_n]$. Now it is easily shown that $[\lambda_1 \mathbf{x}_1 | \lambda_2 \mathbf{x}_2 | \dots | \lambda_n \mathbf{x}_n] = \mathbf{U}\mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_i), 1 \leq i \leq n$. Thus,

$$\mathbf{AU} = \mathbf{U}\mathbf{\Lambda} \quad (4.123)$$

and since \mathbf{U} is invertible (it has rank n)

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}. \quad (4.124)$$

Thus, \mathbf{A} can be decomposed as the product of a diagonal matrix, an invertible matrix and its inverse. In the language of linear algebra [112] it is customary to say that \mathbf{A} is *similar* to the diagonal matrix $\mathbf{\Lambda}$. On the other hand, if (4.124) is true, then \mathbf{A} must have n linearly independent eigenvectors. Although it is known from linear algebra which matrices have property (4.124) the only situation that will concern us is that when \mathbf{A} is symmetric. That symmetric matrices are similar to diagonal matrices is one of the most important facts of matrix algebra and as we will show, this result, the *spectral theorem*, has numerous applications in regression analysis.

Theorem 4.11 (Spectral theorem) *Let \mathbf{A} be an $n \times n$ symmetric matrix, then \mathbf{A} is similar to a diagonal matrix. Moreover, as a consequence of the orthogonality of the eigenvectors of \mathbf{A} , \mathbf{U} in (4.124) can be chosen to be orthogonal so that*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad (4.125)$$

where the columns of \mathbf{U} are orthonormal eigenvectors of \mathbf{A} and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of \mathbf{A} .

Proof. We will show, using an inductive argument, that there exists an orthogonal matrix \mathbf{U} such that $\mathbf{U}^T \mathbf{A} \mathbf{U} = \mathbf{\Lambda}$. From this (4.125) follows and by using the argument immediately following Theorem 4.10 it follows that the columns of \mathbf{U} are eigenvectors of \mathbf{A} and the diagonal elements of $\mathbf{\Lambda}$, the corresponding eigenvalues.

If \mathbf{A} is a 1×1 matrix $[\mathbf{x}]$, then the theorem is trivially true by taking $\mathbf{x}_1 = (1)$. Suppose then that the theorem is true for all $(n-1) \times (n-1)$ symmetric matrices. Now

if \mathbf{A} is $n \times n$, then it has at least one real eigenvalue λ_1 and corresponding eigenvector \mathbf{x}_1 , such that $\|\mathbf{x}_1\| = 1$.

By the Gram-Schmidt process there exist vectors $\mathbf{u}_2, \mathbf{u}_3, \dots, \mathbf{u}_n$ such that $\mathbf{x}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ is an orthonormal basis for \mathbb{R}^n . Let $\mathbf{U}_1 = [\mathbf{x}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_n]$ and observe that $\mathbf{U}_1^T \mathbf{U}_1 = \mathbf{I}_n$ since \mathbf{U}_1 is orthogonal.

Now let

$$\mathbf{B}_1 = \mathbf{U}_1^T \mathbf{A} \mathbf{U}_1, \quad (4.126)$$

then \mathbf{B}_1 has the form

$$\mathbf{B}_1 = \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{A}_2 & \\ \cdot & & & \\ 0 & & & \end{array} \right] \quad (4.127)$$

where \mathbf{A}_2 is an $(n-1) \times (n-1)$ symmetric matrix.

To see this, consider

$$\mathbf{A} \mathbf{U}_1 = [\mathbf{A} \mathbf{x}_1 | \mathbf{A} \mathbf{u}_2 | \dots | \mathbf{A} \mathbf{u}_n] = [\lambda_1 \mathbf{x}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_n] \quad (4.128)$$

where $\mathbf{v}_i = \mathbf{A} \mathbf{u}_i$. Thus,

$$\mathbf{B}_1 = \mathbf{U}_1^T \mathbf{A} \mathbf{U}_1 = [\lambda_1 \mathbf{U}_1^T \mathbf{x}_1 | \mathbf{U}_1^T \mathbf{v}_2 | \dots | \mathbf{U}_1^T \mathbf{v}_n]. \quad (4.129)$$

Since \mathbf{x}_1 is the first column of \mathbf{U}_1 and \mathbf{U}_1 is orthogonal,

$$\mathbf{U}_1^T \mathbf{x}_1 = \left[\begin{array}{c} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle \\ \langle \mathbf{u}_1, \mathbf{x}_1 \rangle \\ \cdot \\ \cdot \\ \cdot \\ \langle \mathbf{u}_n, \mathbf{x}_1 \rangle \end{array} \right] \mathbf{x}_1 = [\langle \mathbf{x}_1, \mathbf{x}_1 \rangle \langle \mathbf{u}_1, \mathbf{x}_1 \rangle \cdots \langle \mathbf{u}_n, \mathbf{x}_1 \rangle]^T = (1, 0, \dots, 0)^T \quad (4.130)$$

so that \mathbf{B}_1 can be partitioned as

$$\mathbf{B}_1 = \left[\begin{array}{c|ccc} \lambda_1 & \alpha_2 & \cdots & \alpha_{n-1} \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{A}_2 & \\ \cdot & & & \\ 0 & & & \end{array} \right]. \quad (4.131)$$

Since \mathbf{A} is symmetric, then $\mathbf{B}_1^T = \mathbf{U}_1^T \mathbf{A}^T \mathbf{U}_1 = \mathbf{U}_1^T \mathbf{A} \mathbf{U}_1 = \mathbf{B}_1$ so that \mathbf{B}_1 is symmetric as well. Thus, from (4.131) $\alpha_2 = \alpha_3 = \dots = \alpha_n = 0$ and \mathbf{A}_2 must be an $(n-1) \times (n-1)$ symmetric matrix. Thus,

$$\mathbf{U}_1^T \mathbf{A} \mathbf{U}_1 = \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{A}_2 & \\ \cdot & & & \\ 0 & & & \end{array} \right]. \quad (4.132)$$

Applying the induction hypothesis to \mathbf{A}_2 , there exists an $(n-1) \times (n-1)$ orthogonal matrix $\hat{\mathbf{U}}_2$ such that $\hat{\mathbf{U}}_2^T \mathbf{A}_2 \hat{\mathbf{U}}_2 = \mathbf{\Lambda}_2$ where $\mathbf{\Lambda}_2$ is diagonal. Now define \mathbf{U}_2 in block form by

$$\mathbf{U}_2 = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \hat{\mathbf{U}}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] \quad (4.133)$$

where \mathbf{U}_2 is easily shown to be orthogonal. Now

$$\begin{aligned} \mathbf{B}_1 \mathbf{U}_2 &= \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{A}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \hat{\mathbf{U}}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] \\ &= \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{A}_2 \hat{\mathbf{U}}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] \end{aligned} \quad (4.134)$$

as may be verified by the block multiplication rules of Section 4.4. Thus,

$$\begin{aligned} \mathbf{U}_2^T \mathbf{B}_1 \mathbf{U}_2 &= \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \hat{\mathbf{U}}_2^T & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{A}_2 \mathbf{U}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] \\ &= \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \hat{\mathbf{U}}_2^T \mathbf{A}_2 \hat{\mathbf{U}}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right] = \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{\Lambda}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right]. \end{aligned} \quad (4.135)$$

Using $\mathbf{B}_1 = \mathbf{U}_1^T \mathbf{A} \mathbf{U}_1$ in (4.135)

$$\begin{aligned} \mathbf{U}_2^T \mathbf{U}_1^T \mathbf{A} \mathbf{U}_1 \mathbf{U}_2 &= (\mathbf{U}_1 \mathbf{U}_2)^T \mathbf{A} (\mathbf{U}_1 \mathbf{U}_2) \\ &= \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \cdot & & & \\ \cdot & & \mathbf{\Lambda}_2 & \\ \cdot & & & \\ \cdot & & & \\ 0 & & & \end{array} \right]. \end{aligned} \quad (4.136)$$

Since $\mathbf{U}_1\mathbf{U}_2$ is an orthogonal matrix, letting $\mathbf{U} = \mathbf{U}_1\mathbf{U}_2$ in (4.136) we find that

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{\Lambda} \quad (4.137)$$

where $\mathbf{\Lambda}$ is a diagonal matrix. This gives, as we observed above, the proof of the theorem. ■

As our first application of the spectral theorem, we give a further decomposition of a symmetric matrix when the matrix is *positive definite*.

Definition 4.4 Let \mathbf{A} an $n \times n$ symmetric matrix, We say that \mathbf{A} is *positive definite* if and only if for every $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq 0$, $\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle > 0$. If only $\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle \geq 0$, $\mathbf{x} \neq 0$, then \mathbf{A} is said to be *positive semidefinite*.

Theorem 4.12 (Properties of positive definite matrices)

- (i) If \mathbf{A} is positive definite, then \mathbf{A} is invertible.
- (ii) \mathbf{A} is positive definite if and only if all the eigenvalues of \mathbf{A} are positive.
- (iii) If \mathbf{A} is positive definite, then there exists a nonsingular matrix \mathbf{R} such that $\mathbf{A} = \mathbf{R}\mathbf{R}^T$.

Proof. (i) Suppose \mathbf{A} is singular. Then there is a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{x} = 0$. Thus, $\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \langle \mathbf{x}, 0 \rangle = 0$ and this contradicts the assumption that \mathbf{A} is positive definite.

(ii) Suppose \mathbf{A} is positive definite and λ is an eigenvalue of \mathbf{A} . Then $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, so that $\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \lambda \langle \mathbf{x}, \mathbf{x} \rangle$. Since $\mathbf{x} \neq 0$, $\lambda = \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle > 0$.

On the other hand, suppose that \mathbf{A} is symmetric with positive eigenvalues. Then by the spectral theorem $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where $\mathbf{\Lambda} = \text{diag}(\lambda_i)$ and $\lambda_i, i = 1, 2, \dots, n$, are the eigenvalues of \mathbf{A} . Then, if $\mathbf{x} \neq 0$,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{\Lambda}\mathbf{U}^T\mathbf{x}, \mathbf{U}^T\mathbf{x} \rangle. \quad (4.138)$$

Let $\mathbf{z} = \mathbf{U}^T\mathbf{x}$, then $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{\Lambda}\mathbf{z}, \mathbf{z} \rangle = \sum_{i=1}^n \lambda_i z_i^2 > 0$, since $\lambda_i > 0$ and $\mathbf{z} \neq 0$.

(iii) From the spectral theorem and (ii), $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where again $\mathbf{\Lambda} = \text{diag}(\lambda_i)$ and $\lambda_i > 0, 1 \leq i \leq n$, are the eigenvalues of \mathbf{A} . Let $\sqrt{\mathbf{\Lambda}} = \text{diag}(\sqrt{\lambda_i})$, then

$$\mathbf{A} = \mathbf{U}\sqrt{\mathbf{\Lambda}}\sqrt{\mathbf{\Lambda}}\mathbf{U}^T = \mathbf{R}\mathbf{R}^T \quad (4.139)$$

where $\mathbf{R} = \mathbf{U}\sqrt{\mathbf{\Lambda}}$. ■

4.6.3 Some Further Applications of the Spectral Theorem

Using the spectral theorem we can establish a number of useful properties of a symmetric matrix \mathbf{A} and its eigenvalues.

Theorem 4.13 If \mathbf{A} is an $n \times n$ symmetric matrix and $\lambda_i, 1 \leq i \leq n$, are the eigenvalues of \mathbf{A} , then

- (i) the eigenvalues of \mathbf{A}^p are λ_i^p ;
- (ii) if \mathbf{A} is invertible, then the eigenvalues of \mathbf{A}^{-1} are $1/\lambda_i, 1 \leq i \leq n$;
- (iii) $\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$;
- (iv) $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i$.

Proof. (i) From the spectral theorem $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where $\mathbf{\Lambda} = \text{diag}(\lambda_i)$. Using the orthogonality of \mathbf{U} it then follows that

$$\mathbf{A}^p = \mathbf{U}\mathbf{\Lambda}^p\mathbf{U}^T. \quad (4.140)$$

Hence, $\mathbf{A}^p\mathbf{U} = \mathbf{U}\mathbf{\Lambda}^p$. Thus the columns of \mathbf{U} are eigenvectors of \mathbf{A}^p with eigenvalues $\lambda_i^p, 1 \leq i \leq n$. Since an $n \times n$ matrix has at most n eigenvalues, all the eigenvalues of \mathbf{A}^p are of the form λ_i^p where λ_i is an eigenvalue of \mathbf{A} .

(ii) This follows as for (i) using the fact that $\mathbf{A}^{-1} = \mathbf{U}^T\mathbf{\Lambda}^{-1}\mathbf{U}$ since \mathbf{A}^{-1} is symmetric and $\mathbf{\Lambda}^{-1} = \text{diag}(1/\lambda_i)$.

(iii) Again using the spectral theorem and property (iv) of the trace we get

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T) = \text{tr}(\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}) = \text{tr}(\mathbf{I}_n\mathbf{\Lambda}) = \text{tr}(\mathbf{\Lambda}) = \sum_{i=1}^n \lambda_i. \quad (4.141)$$

From (4.141) and (ii) it follows that $\text{tr}(\mathbf{A}^{-1}) = \sum_{i=1}^n 1/\lambda_i$.

(iv) Using the fact that the determinant of a product of matrices is the product of the determinants

$$\begin{aligned} \det(\mathbf{A}) &= \det(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T) = \det(\mathbf{U})\det(\mathbf{U}^T)\det(\mathbf{\Lambda}) \\ &= \det(\mathbf{U}\mathbf{U}^T)\det(\mathbf{\Lambda}) = \det(\mathbf{I}_n)\det(\mathbf{\Lambda}) = \prod_{i=1}^n \lambda_i \end{aligned} \quad (4.142)$$

since the determinant of a diagonal matrix is the product of its diagonal elements. ■

Using Theorem 4.12 we prove the important fact that if \mathbf{A} is an orthogonal projection, then

$$\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A}). \quad (4.143)$$

Since $\mathbf{A}^2 = \mathbf{A}$ and \mathbf{A} is symmetric, it follows from Theorem 4.13 that the eigenvalues of \mathbf{A} satisfy $\lambda_i^2 = \lambda_i, 1 \leq i \leq n$. Thus, each λ_i is either 0 or 1. Hence,

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{\Lambda} \quad (4.144)$$

where $\mathbf{\Lambda}$ is a diagonal matrix with p ones and $n - p$ zeros on the diagonal. Since multiplication of a matrix by a nonsingular matrix does not change its rank, the rank of \mathbf{A} is the rank of $\mathbf{\Lambda}$. But the rank of $\mathbf{\Lambda}$ is p since its column space is spanned by the p columns which correspond to the ones on the diagonal. Thus,

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{\Lambda}) = p = \text{tr}(\mathbf{\Lambda}) = \text{tr}(\mathbf{A}). \quad (4.145)$$

4.6.4 Expectation of Quadratic Forms

As an another application of matrix techniques in this Chapter we calculate the expected value of a quadratic form of a random vector. Such expectations will be quite useful in the following Chapter.

Let (Y_1, Y_2, \dots, Y_n) be n random variables. Collectively we will call them a *random vector* \mathbf{Y} ,

Definition 4.5 Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ be a random vector. The *mean vector* of \mathbf{Y} , $E(\mathbf{Y})$ is the vector

$$E(\mathbf{Y}) = [E(Y_1), E(Y_2), \dots, E(Y_n)]^T \quad (4.146)$$

provided the expectations exist. Also, the *variance-covariance matrix* of \mathbf{Y} , $\Sigma(\mathbf{Y})$ is the $n \times n$ matrix

$$\Sigma(\mathbf{Y}) = [\text{Cov}(Y_i, Y_j)], 1 \leq i, j \leq n. \quad (4.147)$$

Note: The diagonal elements of $\Sigma(\mathbf{Y})$ are $\text{Cov}(Y_i, Y_i) = \text{Var}(Y_i)$.

Definition 4.6 Let $\mathbf{A} = [a_{ij}], i, j = 1, 2, \dots, n$ be an $n \times n$ symmetric matrix, and let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ be an $n \times 1$ random vector. If $Z = \langle \mathbf{Y}, \mathbf{A}\mathbf{Y} \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{ij} Y_i Y_j$, we say that Z is a *quadratic form* in the Y_i 's.

Theorem 4.14 Let $Z = \langle \mathbf{Y}, \mathbf{A}\mathbf{Y} \rangle$ be a quadratic form and assume that $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\Sigma(\mathbf{Y}) = \Sigma = [\sigma_{ij}], i, j = 1, 2, \dots, n$. Then

$$E(Z) = \text{tr}(\mathbf{A}\Sigma) + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}. \quad (4.148)$$

Proof. $E(Z) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} E(Y_i Y_j)$. Since $\text{Cov}(Y_i Y_j) = E(Y_i Y_j) - E(Y_i) E(Y_j) = E(Y_i Y_j) - \mu_i \mu_j$, $E(Y_i Y_j) = \text{Cov}(Y_i Y_j) + \mu_i \mu_j = \sigma_{ij} + \mu_i \mu_j$. Thus,

$$E(Z) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \text{Cov}(Y_i Y_j) + \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mu_i \mu_j. \quad (4.149)$$

Now using the definition of the trace and the fact that $\text{tr}(\mathbf{A}\Sigma) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \sigma_{ij}$, we have $E(Z) = \text{tr}(\mathbf{A}\Sigma) + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}$. ■

We now turn our attention to the application of some of these matrix results to some basic results in probability theory, the first being a short discussion of the multivariate normal distribution.

4.7 The Multivariate Normal Distribution

4.7.1 The Nondegenerate Case

Definition 4.6 We say that a random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ has a *multivariate normal density* $f_{\mathbf{Y}}(\mathbf{y})$ if and only if

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma}} \exp \left[-\frac{1}{2} \langle \mathbf{y} - \boldsymbol{\mu}, \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})^T \rangle \right], \mathbf{y} \in \mathbb{R}^n, \quad (4.150)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}$ is a positive-definite symmetric matrix. It is customary in this case to say that \mathbf{Y} has a *nondegenerate multivariate normal distribution*. In this case we write

$$\mathbf{Y} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (4.151)$$

Theorem 4.15 (i) Let \mathbf{Y} have a joint multivariate normal density. Then $f_{\mathbf{Y}}(\mathbf{y})$ as given by (4.150) is a density. That is, $f_{\mathbf{Y}}(\mathbf{y}) \geq 0$ and $\int_{\mathbb{R}^n} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} = 1$.

(ii) \mathbf{Y} has a joint multivariate normal density if and only if $\mathbf{Y} = \mathbf{AZ} + \boldsymbol{\mu}$, where \mathbf{A} is $n \times n$ and nonsingular and $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)^T$ is a vector of n independent $N(0, 1)$ random variables; i.e., $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n)$.

(iii) Let $E(\mathbf{Y}) = [E(Y_1), E(Y_2), \dots, E(Y_n)]^T$ be the mean vector of \mathbf{Y} and $\boldsymbol{\Sigma}(\mathbf{Y}) = [\text{Cov}(Y_i, Y_j)]$, $1 \leq i, j \leq n$, be the variance-covariance matrix of \mathbf{Y} . Then, $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}(\mathbf{Y}) = \boldsymbol{\Sigma}$.

(iv) The random variables Y_i , $1 \leq i \leq n$, are independent if and only if they are uncorrelated; i.e., $[\text{Cov}(Y_i, Y_j)] = \text{diag}(\sigma_i^2)$, where $\sigma_i^2 = \text{Var}(Y_i)$.

(v) The moment generating function of \mathbf{Y} is given by

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp(\langle \mathbf{t}, \boldsymbol{\mu} \rangle) \exp(\langle \mathbf{t}, \boldsymbol{\Sigma} \mathbf{t} \rangle / 2) \quad (4.152)$$

where $\mathbf{t} = (t_1, t_2, \dots, t_n)^T \in \mathbb{R}^n$.

(vi) If $\mathbf{Y} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n)$ and \mathbf{U} is an orthogonal matrix, then $\mathbf{UY} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n)$.

Note: Theorem 4.15 gives the main properties of jointly distributed normal random variables that we will need in this text. A more comprehensive discussion may be found in references such as [40, 63].

Before proving Theorem 4.15 we prove a number of additional properties of random vectors.

Theorem 4.16 Let \mathbf{Y} be a random n -vector and let \mathbf{A} be an $m \times n$ matrix. Then, if \mathbf{Z} is a random m -vector

(i) $E(\mathbf{AY} + \mathbf{Z}) = \mathbf{AE}(\mathbf{Y}) + E(\mathbf{Z})$.

(ii) $\boldsymbol{\Sigma}(\mathbf{Y} + \mathbf{b}) = \boldsymbol{\Sigma}(\mathbf{Y})$, where \mathbf{b} is a constant n -vector.

(iii) $\boldsymbol{\Sigma}(\mathbf{AY}) = \mathbf{A}\boldsymbol{\Sigma}(\mathbf{Y})\mathbf{A}^T$.

Proof. (i) Now the i -th component of $\mathbf{AY} + \mathbf{Z}$ is $\sum_{j=1}^n a_{ij}Y_j + Z_i$ where $\mathbf{A} = [a_{ij}]$, $1 \leq i \leq m$, $1 \leq j \leq n$. By linearity of expectation $E\left(\sum_{j=1}^n a_{ij}Y_j + Z_i\right) = \sum_{j=1}^n a_{ij}E(Y_j) + E(Z_i)$, $1 \leq i \leq m$. But this is the i -th element of $\mathbf{AE}(\mathbf{Y}) + E(\mathbf{Z})$.

(ii) The ij -th element of $\Sigma(\mathbf{Y} + \mathbf{b})$ is given by

$$\begin{aligned}
 \text{Cov}(Y_i + b_i, Y_j + b_j) &= E[(Y_i + b_i)(Y_j + b_j)] - E(Y_i + b_i)E(Y_j + b_j) \\
 &= E[(Y_i Y_j) + b_j E(Y_i) + b_i E(Y_j)] + b_i b_j \\
 &\quad - [E(Y_i) + b_i][E(Y_j) + b_j] \\
 &= E(Y_i Y_j) + b_j E(Y_i) + b_i E(Y_j) + b_i b_j \\
 &\quad - E(Y_i)E(Y_j) - b_j E(Y_i) - b_i E(Y_j) - b_i b_j \\
 &= E(Y_i Y_j) - E(Y_i)E(Y_j) = \text{Cov}(Y_i, Y_j)
 \end{aligned} \tag{4.153}$$

which is the ij -th element of $\Sigma(\mathbf{Y})$.

(iii) Now the i -th element of $\mathbf{X} = \mathbf{A}\mathbf{Y}$ is $\sum_{j=1}^n a_{ij}Y_j$. Hence, using (2.74)

$$\begin{aligned}
 \text{Cov}(X_i, X_j) &= \sum_{l=1}^n \sum_{k=1}^n a_{jk} a_{il} \text{Cov}(Y_k, Y_l) \\
 &= \sum_{l=1}^n \sum_{k=1}^n a_{jk} a_{il} \sigma_{kl}.
 \end{aligned} \tag{4.154}$$

Also,

$$(\mathbf{A}\Sigma)_{il} = \sum_{k=1}^n a_{ik} \sigma_{kl} \equiv b_{il}. \tag{4.155}$$

Denoting the lj -th element of \mathbf{A}^T by $c_{lj} = a_{jl}$ the ij -th element of $\mathbf{A}\Sigma\mathbf{A}^T$ is given by

$$\begin{aligned}
 \sum_{l=1}^n b_{il} c_{lj} &= \sum_{l=1}^n \left(\sum_{k=1}^n a_{ik} \sigma_{kl} \right) a_{jl} = \sum_{l=1}^n \sum_{k=1}^n a_{ik} a_{jl} \sigma_{kl} \\
 &= \text{Cov}(X_j, X_i) = \text{Cov}(X_i, X_j).
 \end{aligned} \tag{4.156}$$

■

Proof of Theorem 4.15. (i) Obviously $f_{\mathbf{Y}}(\mathbf{y}) \geq 0$, so to show that $f_{\mathbf{Y}}(\mathbf{y})$ is a density, it suffices to show that it integrates to one. To do this use Theorem 4.12 to write $\Sigma = \mathbf{R}\mathbf{R}^T$. Then,

$$\begin{aligned}
 \langle \mathbf{y} - \boldsymbol{\mu}, \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}) \rangle &= \left\langle \mathbf{y} - \boldsymbol{\mu}, (\mathbf{R}\mathbf{R}^T)^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right\rangle \\
 &= \langle \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \rangle
 \end{aligned} \tag{4.157}$$

and

$$\sqrt{\det \Sigma} = \sqrt{\det (\mathbf{R}\mathbf{R}^T)} = |\det \mathbf{R}|. \tag{4.158}$$

Let $\mathbf{z} = \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and using the change of variables formula for multiple integrals [40]

$$\begin{aligned}
 \int_{\mathbb{R}^n} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} &= \int_{\mathbb{R}^n} f_{\mathbf{Y}}(\mathbf{R}\mathbf{z} + \boldsymbol{\mu}) |\det \mathbf{R}| d\mathbf{z} \\
 &= \int_{\mathbb{R}^n} \frac{\exp[-\langle \mathbf{z}, \mathbf{z} \rangle / 2]}{(2\pi)^{n/2} |\det \mathbf{R}|} |\det \mathbf{R}| d\mathbf{z} \\
 &= \int_{\mathbb{R}^n} \frac{\exp[-\langle \mathbf{z}, \mathbf{z} \rangle / 2]}{(2\pi)^{n/2}} d\mathbf{z}.
 \end{aligned} \tag{4.159}$$

But $\langle \mathbf{z}, \mathbf{z} \rangle = \sum_{i=1}^n z_i^2$, so that

$$\begin{aligned} \int_{\mathbb{R}^n} \frac{\exp[-\langle \mathbf{z}, \mathbf{z} \rangle / 2]}{(2\pi)^{n/2}} d\mathbf{z} &= \int_{\mathbb{R}^n} \prod_{i=1}^n \left[\frac{\exp(-z_i^2/2)}{(2\pi)^{1/2}} \right] dz_i \\ &= \prod_{i=1}^n \int_{-\infty}^{\infty} \left[\frac{\exp(-z_i^2/2)}{(2\pi)^{1/2}} \right] dz_i = 1, \end{aligned} \quad (4.160)$$

since each of the integrals in (4.160) is the integral of a standard normal density.

(ii) Making use of arguments similar to those in (i) we find that $\mathbf{Y} = \mathbf{RZ} + \boldsymbol{\mu}$ where $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{R}^T$ and $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n)$. The details are left as an exercise.

(iii) From (ii) $E(\mathbf{Y}) = E(\mathbf{RZ} + \boldsymbol{\mu}) = \mathbf{R}E(\mathbf{Z}) + E(\boldsymbol{\mu}) = \boldsymbol{\mu}$, since $E(\mathbf{Z}) = \mathbf{0}$. Also from Theorem 4.14,

$$\begin{aligned} \boldsymbol{\Sigma}(\mathbf{Y}) &= \boldsymbol{\Sigma}(\mathbf{RZ} + \boldsymbol{\mu}) = \boldsymbol{\Sigma}(\mathbf{RZ}) \\ &= \mathbf{R}\boldsymbol{\Sigma}(\mathbf{Z})\mathbf{R}^T = \mathbf{R}\mathbf{I}_n\mathbf{R}^T = \mathbf{R}\mathbf{R}^T = \boldsymbol{\Sigma}. \end{aligned} \quad (4.161)$$

(iv) Suppose that $(Y_1, Y_2, \dots, Y_n)^T$ are $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and are independent. Then

$$\text{Cov}(Y_i, Y_j) = \begin{cases} \sigma_i^2, & i = j \\ 0, & i \neq j. \end{cases} \quad (4.162)$$

On the other hand, since $\boldsymbol{\Sigma}$ is the variance-covariance matrix of \mathbf{Y} and $Y_i, 1 \leq i \leq n$, are uncorrelated, then $\boldsymbol{\Sigma} = \text{diag}(\sigma_i^2)$, $\boldsymbol{\Sigma}^{-1} = \text{diag}(1/\sigma_i^2)$ and

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{\exp[\langle \mathbf{y} - \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \rangle] / 2}{(2\pi)^{n/2} \prod_{i=1}^n \sigma_i} \\ &= \prod_{i=1}^n \left\{ \frac{\exp[-(y_i - \mu_i)^2 / 2\sigma_i^2]}{(2\pi)^{1/2} \sigma_i} \right\}. \end{aligned} \quad (4.163)$$

Since the joint density of \mathbf{Y} factors as the product n $N(\mu_i, \sigma_i^2)$ densities, it follows that $Y_i, 1 \leq i \leq n$, are independent $N(\mu_i, \sigma_i^2)$ random variables.

(v) By definition, the moment generating function of \mathbf{Y} is

$$M_{\mathbf{Y}}(\mathbf{t}) = E[e^{\langle \mathbf{t}, \mathbf{Y} \rangle}]. \quad (4.164)$$

By (iii) $\mathbf{Y} = \mathbf{RZ} + \boldsymbol{\mu}$ where $\mathbf{R}\mathbf{R}^T = \boldsymbol{\Sigma}$ so that

$$\langle \mathbf{t}, \mathbf{Y} \rangle = \langle \mathbf{t}, \mathbf{RZ} \rangle + \langle \mathbf{t}, \boldsymbol{\mu} \rangle = \langle \mathbf{R}^T \mathbf{t}, \mathbf{Z} \rangle + \langle \mathbf{t}, \boldsymbol{\mu} \rangle = \langle \mathbf{s}, \mathbf{Z} \rangle + \langle \mathbf{t}, \boldsymbol{\mu} \rangle \quad (4.165)$$

where $\mathbf{s} = \mathbf{R}^T \mathbf{t}$. Thus,

$$M_{\mathbf{Y}}(\mathbf{t}) = e^{\langle \mathbf{t}, \boldsymbol{\mu} \rangle} E(e^{\langle \mathbf{s}, \mathbf{Z} \rangle}). \quad (4.166)$$

But $\langle \mathbf{s}, \mathbf{Z} \rangle = \sum_{i=1}^n s_i Z_i$ so that $E(e^{\langle \mathbf{s}, \mathbf{Z} \rangle}) = \prod_{i=1}^n E(e^{s_i Z_i})$ where $Z_i, 1 \leq i \leq n$, are $N(0, 1)$ random variables.

From Section 2.7

$$\begin{aligned}
 E\left(e^{\langle \mathbf{s}, \mathbf{Z} \rangle}\right) &= \prod_{i=1}^n \exp(s_i^2/2) = \exp\left(\sum_{i=1}^n s_i^2/2\right) \\
 &= \exp(\langle \mathbf{s}, \mathbf{s} \rangle / 2) = \exp(\langle \mathbf{R}^T \mathbf{t}, \mathbf{R}^T \mathbf{t} \rangle / 2) \\
 &= \exp(\langle \mathbf{t}, \mathbf{R} \mathbf{R}^T \mathbf{t} \rangle / 2) = \exp(\langle \mathbf{t}, \mathbf{\Sigma} \mathbf{t} \rangle / 2).
 \end{aligned} \tag{4.167}$$

Combining (4.166) and (4.167) we find that

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp(\langle \mathbf{t}, \boldsymbol{\mu} \rangle) \exp(\langle \mathbf{t}, \mathbf{\Sigma} \mathbf{t} \rangle / 2). \tag{4.168}$$

(vi) We establish (vi) by calculating the moment generating function of $\mathbf{W} = \mathbf{U}\mathbf{Y}$. From (v) $M_{\mathbf{Y}}(\mathbf{t}) = e^{\langle \mathbf{t}, \mathbf{t} \rangle / 2}$, since $\mathbf{\Sigma}(\mathbf{Y}) = \mathbf{I}_n$. Thus,

$$\begin{aligned}
 M_{\mathbf{W}}(\mathbf{t}) &= E\left(e^{\langle \mathbf{t}, \mathbf{U}\mathbf{Y} \rangle}\right) = E\left(e^{\langle \mathbf{U}^T \mathbf{t}, \mathbf{Y} \rangle}\right) \\
 &= e^{\langle \mathbf{U}^T \mathbf{t}, \mathbf{U}^T \mathbf{t} \rangle / 2} = e^{\langle \mathbf{t}, \mathbf{U} \mathbf{U}^T \mathbf{t} \rangle / 2} = e^{\langle \mathbf{t}, \mathbf{t} \rangle / 2},
 \end{aligned} \tag{4.169}$$

since $\mathbf{U} \mathbf{U}^T = \mathbf{I}_n$. Thus \mathbf{W} has the same moment generating function as \mathbf{Y} and so \mathbf{W} has the same joint distribution as \mathbf{Y} . Thus $\mathbf{W} = (W_1, W_2, \dots, W_n)^T$ are independent $N(0, 1)$ random variables. ■

4.7.2 The Degenerate Multivariate Normal Distribution

In order to develop the distribution theory associated with multiple regression models with normal errors, it is necessary to generalize the multivariate normal distribution to the case where $\mathbf{\Sigma}$ is singular. In this case \mathbf{Y} will not have a density so (4.150) will not apply. However, if we use the well known fact from probability theory that the distribution of many random variables may be characterized in terms of their moment generating functions, then we can use (4.152) to define joint multivariate normal distributions with singular variance-covariance matrices. Such distributions are usually said to be *degenerate* since they do not have densities.

Definition 4.7 Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ be an n -vector of random variables. We say that \mathbf{Y} has a *degenerate multivariate normal distribution* if and only if the joint moment generating function of \mathbf{Y} is given by

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp(\langle \mathbf{t}, \boldsymbol{\mu} \rangle + \langle \mathbf{t}, \mathbf{\Sigma} \mathbf{t} \rangle / 2). \tag{4.170}$$

where $\mathbf{\Sigma}$ is a positive semidefinite matrix.

As before, we write $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Theorem 4.17 Assume that \mathbf{Y} has a degenerate $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution. Then,

- (i) $\mathbf{Y} = (\mathbf{R}\mathbf{Z} + \boldsymbol{\mu})$, where $\mathbf{\Sigma} = \mathbf{R}\mathbf{R}^T$ and \mathbf{Z} is $N(\mathbf{0}, \mathbf{I}_n)$.
- (ii) $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\mathbf{\Sigma}(\mathbf{Y}) = \mathbf{\Sigma}$.
- (iii) If $\text{Var}(Y_i) = \sigma^2 > 0$, then Y_i is $N(\mu_i, \sigma_i^2)$. Otherwise Y_i is a degenerate random variable with $P\{Y_i = \mu_i\} = 1$.

(iv) $Y_i, 1 \leq i \leq n$, are independent random variables if and only if they are uncorrelated.

Proof. (i) To prove this we find the moment generating function of \mathbf{Y} . Straight-forward manipulations as in Theorem 4.15 show that the moment generating function of $\mathbf{RZ} + \boldsymbol{\mu}$ is given by (4.170). Thus, $\mathbf{RZ} + \boldsymbol{\mu}$ and \mathbf{Y} have the same distribution.

(ii) This follows as in Theorem 4.15.

(iii) The marginal distribution of Y_i is given by setting $t_j = 0, j \neq i$, in (4.170). In this case, $\langle \mathbf{t}, \boldsymbol{\Sigma} \mathbf{t} \rangle = \sigma_i^2 t_i^2$ where σ_i^2 is the i -th diagonal element of $\boldsymbol{\Sigma}$ and $\langle \mathbf{t}, \boldsymbol{\mu} \rangle = \mu_i, 1 \leq i \leq n$. Thus,

$$M_{Y_i}(t_i) = e^{t_i \mu_i} e^{\sigma_i^2 t_i^2 / 2}. \quad (4.171)$$

If $\sigma_i^2 > 0$, then Y_i is $N(\mu_i, \sigma_i^2)$. On the other hand, if $\sigma_i^2 = 0$ then

$$M_{Y_i}(t_i) = e^{t_i \mu_i} \quad (4.172)$$

and this is the moment generating function of a degenerate random variable D_i with $P\{D_i = \mu_i\} = 1$. Then Y_i has the same distribution as D_i .

(iv) Using (iii) this follows as in Theorem 4.15. ■

4.8 Solving Systems of Equations

As we have seen in Chapter 3, the computation of parameter estimates in the simple linear regression model requires the solution of two linear equations in two unknowns. For the models to be treated in subsequent chapters involving $m > 2$ parameters this will require the solution of m equations in m unknowns, which can be tedious, even for moderate values of m . Although most modern statistical packages treat this computation as a “black box”, we feel it is useful for students to see how this can be done and perhaps write their own programs. (It is the authors’ experience that these black boxes do not always work for all problems and one may sometimes have to do it himself/herself.)

As we observed at the beginning of the Chapter, the set of m equations in m unknowns (x_1, x_2, \dots, x_m)

$$\sum_{j=1}^m a_{ij} x_j = b_i, \quad 1 \leq i \leq m, \quad (4.173)$$

can be written in matrix-vector form as

$$\mathbf{A} \mathbf{x} = \mathbf{b} \quad (4.174)$$

where $\mathbf{A} = [a_{ij}], 1 \leq i, j \leq m$, is the *coefficient matrix*, $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$ and $\mathbf{b} = (b_1, b_2, \dots, b_m)^T$. Now (4.174) has a unique solution if and only if \mathbf{A} is invertible and then

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b} \quad (4.175)$$

solves (4.174). Thus, if we can compute \mathbf{A}^{-1} , then (4.175) gives as the required solution \mathbf{x} . Unfortunately, computing \mathbf{A}^{-1} is usually a more difficult task than solving (4.174) directly, so one does not usually proceed in this fashion numerically. For small values of m it is possible to use Cramer’s rule [112], but again, this is usually not computationally desirable for most practical problems.

Historically, the most widely used technique for solving (4.174) has been *Gaussian elimination* which is a generalization of the well known elimination method taught in high school algebra. This method is still widely used, for both statistical and nonstatistical problems [112, 120] and is generally quite efficient and accurate for moderate values of m , provided \mathbf{A} is well-conditioned (see Section 4.9 for a definition) and can be readily programmed. Because the coefficient matrices in regression analysis are of the form

$$\mathbf{A} = \mathbf{X}^T \mathbf{X}, \quad (\mathbf{X} \text{ is } n \times m) \quad (4.176)$$

they will be positive definite if \mathbf{X} has full rank. In this case it is often useful to exploit this property to produce more efficient and stable algorithms and we shall discuss several of these as well.

4.8.1 Gaussian Elimination

With Gaussian elimination, as well as many other techniques, the goal is to reduce the system (4.174) to an equivalent triangular system which can be solved by *backward* or *forward* substitution. For example, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}, \quad (4.177)$$

then (4.174) becomes

$$\begin{aligned} (1) \quad & a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ (2) \quad & a_{22}x_2 + a_{23}x_3 = b_2, \\ (3) \quad & a_{33}x_3 = b_3. \end{aligned}$$

One then proceeds by solving (3) for x_3 substituting in (2) solving for x_2 and finally substituting x_3, x_2 into (1) to get x_1 . Clearly, this process can be generalized when \mathbf{A} is upper triangular

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ 0 & a_{22} & \cdots & a_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & a_{mm} \end{bmatrix}. \quad (4.178)$$

In Gaussian elimination we first reduce \mathbf{A} in (4.174) to the form in (4.178) and then the resulting system is solved by back substitution. The reduction is carried out by successively eliminating variables in all equations below the first. For simplicity, we illustrate this process for $m = 3$.

Hence, we consider the following equations in three unknowns,

$$\begin{aligned} (1) \quad & a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ (2) \quad & a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2, \\ (3) \quad & a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3. \end{aligned}$$

We begin by trying to eliminate the variable x_1 in equations (2) and (3). We do this by adding suitable multiples of (1) to (2), and then to (3). To eliminate x_1 in (2) we must make $a_{21} = 0$. Thus we need to determine c so that $a_{21} + ca_{11} = 0$. If $a_{11} \neq 0$ (which we assume) then $c = -a_{21}/a_{11}$. Thus adding $c \times (1)$ to (2) the second equation becomes

$$(2') \quad 0x_1 + \left[a_{22} - \left(\frac{a_{21}}{a_{11}} \right) a_{12} \right] x_2 + \left[a_{23} - \left(\frac{a_{21}}{a_{11}} \right) a_{13} \right] x_3 = b_2 - \left(\frac{a_{21}}{a_{11}} \right) b_1. \quad (4.179)$$

Similarly, eliminating x_1 from (3) gives

$$(3') \quad 0x_1 + \left[a_{32} - \left(\frac{a_{31}}{a_{11}} \right) a_{12} \right] x_2 + \left[a_{33} - \left(\frac{a_{31}}{a_{11}} \right) a_{13} \right] x_3 = b_3 - \left(\frac{a_{31}}{a_{11}} \right) b_1. \quad (4.180)$$

The transformed equations now take the form

$$\begin{aligned} (1') \quad & a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ (2') \quad & a'_{22}x_2 + a'_{23}x_3 = b'_2, \\ (3') \quad & a'_{32}x_2 + a'_{33}x_3 = b'_3, \end{aligned}$$

where a'_{22} , a'_{23} etc. are determined as in (4.179)-(4.180). This procedure may now be repeated, eliminating x_2 from the last equation by adding the multiple $-a'_{32}/a'_{22}$ (2') to (3'). This results in the system

$$\begin{aligned} (1'') \quad & a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ (2'') \quad & a'_{22}x_2 + a'_{23}x_3 = b'_2, \\ (3'') \quad & a''_{33}x_3 = b'_3. \end{aligned}$$

If the original equations have a unique solution for every right hand side, this procedure always works theoretically, subject to the following proviso. Note that at each stage we had to divide by the *pivot elements* on the diagonal. So these elements need to be non-zero. This may not always be the case, but can always be accomplished by a row interchange because of the unique solvability of the equations. Thus, an additional ingredient of the algorithm requires one to provide a mechanism to search for a non-zero pivot element at each stage. One of the nice features about the least squares equations is that because $\mathbf{X}^T\mathbf{X}$ is positive definite, this guarantees that successive pivot elements are non-zero. However, in many problems some of the pivots can become very small and then dividing by these can produce substantial *round-off error*. Hence, one often uses a *partial pivoting* strategy of searching the leading elements in the pivot column below the current pivot a_{ii} for the largest element (in absolute value) and then interchanging that row with the i -th.

Example 4.2 To illustrate some of these ideas we solve the following equations by Gaussian elimination

$$\begin{aligned} (1) \quad & 2x_1 + x_2 + x_3 = 4, \\ (2) \quad & x_1 + x_2 + x_3 = 4, \\ (3) \quad & 3x_1 + 2x_2 + x_3 = 6. \end{aligned}$$

Eliminating x_1 in (2) and (3) gives

$$\begin{aligned}
(1') \quad & 2x_1 + x_2 + x_3 = 4, \\
(2') \quad & x_2/2 + x_3/2 = 2, \\
(3') \quad & x_2/2 - x_3/2 = 0.
\end{aligned}$$

Then, eliminating x_2 in (3') we get

$$\begin{aligned}
(1'') \quad & 2x_1 + x_2 + x_3 = 4, \\
(2'') \quad & x_2/2 + x_3/2 = 2, \\
(3'') \quad & -x_3 = -2.
\end{aligned}$$

Solving these by back substitution gives: $x_3 = 2$, $x_2/2 = 2 - 1$ or $x_2 = 2$; and $2x_1 = 4 - 2 - 2 = 0$ or $x_1 = 0$. Hence, the solution is: $x_1 = 0, x_2 = 2, x_3 = 2$.

Example 4.3 Solve the following system of equations by Gaussian elimination.

$$\begin{aligned}
(1) \quad & x_2 + x_3 = 2, \\
(2) \quad & x_1 + x_2 = 2, \\
(3) \quad & x_1 + x_3 = 2,
\end{aligned}$$

Here a_{11} is zero, so we begin by interchanging (1) and (2) to get

$$\begin{aligned}
(1') \quad & x_1 + x_2 = 2, \\
(2') \quad & x_2 + x_3 = 2, \\
(3') \quad & x_1 + x_3 = 2.
\end{aligned}$$

Eliminating x_1 in (3') by subtracting (1') from (3') gives

$$\begin{aligned}
(1'') \quad & x_1 + x_2 = 2, \\
(2'') \quad & x_2 + x_3 = 2, \\
(3'') \quad & -x_2 + x_3 = 0.
\end{aligned}$$

Adding (2'') to (3'') gives: $2x_3 = 2$, so $x_3 = 1$; and by back substitution into (2'') gives $x_2 = 1$; and again putting x_2 into (1'') we obtain $x_1 = 1$. Hence, the solution is: $x_1 = 1, x_2 = 1, x_3 = 1$.

For numerical purposes manipulation with the unknowns $\{x_i\}$ are superfluous. All the calculations can be done by performing the successive operations on the *augmented coefficient matrix*.

$$[\mathbf{A}|\mathbf{b}] = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ a_{12} & \cdots & \cdot & b_2 \\ \cdot & & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ a_{m1} & \cdots & a_{mm} & b_m \end{array} \right] \quad (4.181)$$

The elimination steps are performed on $[\mathbf{A}|\mathbf{b}]$ until it becomes (after $m - 1$ steps)

$$\left[\begin{array}{ccc|c} a_{11}^{(m-1)} & \cdots & a_{1n}^{(m-1)} & b_1^{(m-1)} \\ 0 & \cdots & \cdot & b_2^{(m-1)} \\ \cdot & & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ 0 & 0 & a_{mm}^{(m-1)} & b_m^{(m-1)} \end{array} \right] \quad (4.182)$$

and then the resulting equations are solved by back substitution.

An important consequence of the Gaussian elimination process is that it is equivalent to *factoring* \mathbf{A} as [112]

$$\mathbf{A} = \mathbf{LU} \quad (4.183)$$

where \mathbf{L} is a lower triangular matrix formed from the multipliers used to do the elimination and \mathbf{U} is the coefficient matrix of the final upper triangular system. Such matrix factorizations are fundamental in modern numerical analysis.

Example 4.4 If we consider the system in Example 4.3 we find the matrix of multipliers is given by

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 3/2 & 1 & 1 \end{bmatrix}$$

and

$$\mathbf{U} = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & -1 \end{bmatrix}.$$

Straightforward matrix multiplication shows that $\mathbf{A} = \mathbf{LU}$.

Computing \mathbf{A}^{-1}

Because regression analysis requires not only parameter estimates but standard errors and correlations, this requires obtaining \mathbf{A}^{-1} as well. This can also be done using Gaussian elimination or equivalently the \mathbf{LU} factorization.

If \mathbf{A} is an $m \times m$ invertible matrix, then \mathbf{A}^{-1} is the unique matrix solving

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_m. \quad (4.184)$$

If we write

$$\mathbf{A}^{-1} = [\mathbf{x}_1 | \mathbf{x}_2 | \dots | \mathbf{x}_m] \quad (4.185)$$

then (4.184) becomes

$$[\mathbf{A}\mathbf{x}_1 | \mathbf{A}\mathbf{x}_2 | \dots | \mathbf{A}\mathbf{x}_m] = \mathbf{I}_m. \quad (4.186)$$

Since the i -th column of \mathbf{I}_m is the i -th canonical basis element \mathbf{e}_i (4.186) is equivalent to solving the m systems of equations

$$\mathbf{A}\mathbf{x} = \mathbf{e}_i, \quad 1 \leq i \leq m. \quad (4.187)$$

Again, these systems can be solved by Gaussian elimination by performing elimination on the matrix

$$[\mathbf{A} | \mathbf{I}_m]. \quad (4.188)$$

Equivalently, if one has the \mathbf{LU} decomposition, then (4.187) becomes

$$\mathbf{LU}\mathbf{x}_i = \mathbf{e}_i, \quad 1 \leq i \leq m. \quad (4.189)$$

These can be solved by setting $\mathbf{z}_i = \mathbf{U}\mathbf{x}_i$, $1 \leq i \leq m$ and solving for \mathbf{z}_i from

$$\mathbf{L}\mathbf{z}_i = \mathbf{e}_i \quad (4.190)$$

and then \mathbf{x}_i is obtained from

$$\mathbf{U}\mathbf{x}_i = \mathbf{z}_i. \quad (4.191)$$

This is easily done by forward substitution in (4.189) and backward substitution in (4.190).

4.8.2 Cholesky Factorization

As we observed above, Gaussian elimination does not use any properties of \mathbf{A} other than its invertibility. When \mathbf{A} is positive definite, other factorizations may be more convenient and numerically less prone to round-off error. One of the most important of these is the *Cholesky factorization*

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T \quad (4.192)$$

where \mathbf{L} is a lower triangular matrix. This factorization can be obtained by assuming \mathbf{L} in the form

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdot \\ \cdot & \cdot & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{bmatrix} \quad (4.193)$$

and then using (4.192) to solve explicitly for l_{ij} . We illustrate this process for a 3×3 matrix. (Note: Because \mathbf{A} is symmetric we only need the lower triangle of $\mathbf{L}\mathbf{L}^T$.)

Now,

$$\begin{aligned} \mathbf{L}\mathbf{L}^T &= \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} \\ &= \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{31}l_{11} & l_{31}l_{21} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix} \end{aligned} \quad (4.194)$$

Equating elements in (4.194) gives $l_{11}^2 = a_{11}$ so that $l_{11} = \sqrt{a_{11}}$. (This makes sense since the diagonal elements of a positive definite matrix are positive. In fact, $a_{ii} = \langle \mathbf{e}_i, \mathbf{A}\mathbf{e}_i \rangle > 0$.) Equating elements in the second row gives

$$l_{21}l_{11} = a_{21} \text{ and } l_{21}^2 + l_{22}^2 = a_{22}. \quad (4.195)$$

Thus, $l_{21} = a_{21}/l_{11}$, $l_{22}^2 = a_{22} - a_{21}^2/a_{11} = (a_{11}a_{22} - a_{21}^2)/a_{11}$. It follows from the positive definiteness of \mathbf{A} that $a_{11}a_{22} - a_{21}^2 > 0$ so that $l_{22} = \sqrt{(a_{11}a_{22} - a_{21}^2)/a_{11}}$.

Last, equating elements the third row gives

$$l_{31}l_{11} = a_{31}, \quad l_{31}l_{21} + l_{22}l_{32} = a_{32} \text{ and } l_{31}^2 + l_{32}^2 + l_{33}^2 = a_{33} \quad (4.196)$$

and these can be solved successively for l_{31} , l_{32} and l_{33} .

For a general $m \times m$ positive definite symmetric matrix this process can be extended giving the i -th row l_{ij} , $j = 1, 2, \dots, i-1$ as [112]

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}}, \quad j = 1, 2, \dots, i-1. \quad (4.197)$$

As with the \mathbf{LU} factorization, the Cholesky factorization can be used to solve (4.174). In this case

$$\mathbf{L}\mathbf{L}^T \mathbf{x} = \mathbf{b} \quad (4.198)$$

and setting $\mathbf{z} = \mathbf{L}^T \mathbf{x}$, $\mathbf{Lz} = \mathbf{b}$. As before, \mathbf{z} can be obtained by forward substitution and \mathbf{x} by back substitution. The inverse of \mathbf{A} can be found as for the \mathbf{LU} decomposition by solving

$$\mathbf{LL}^T \mathbf{x}_i = \mathbf{e}_i, \quad 1 \leq i \leq m. \quad (4.199)$$

The Cholesky factorization is important not only for solving systems of equations, it may be needed to determine other quantities as well. For example, in many statistical calculations one requires the determinant of \mathbf{A} . The Cholesky factorization gives a convenient way of doing this. In fact, $\det(\mathbf{A}) = \det(\mathbf{LL}^T) = \det^2(\mathbf{L})$ since $\det(\mathbf{L}) = \det(\mathbf{L}^T)$. But the determinant of a triangular matrix is the product of its diagonal elements so that $\det^2(\mathbf{L}) = \prod_{i=1}^m l_{ii}^2 > 0$. Historically, $\det(\mathbf{A})$ was often used as a measure of conditioning of \mathbf{A} ; small values indicating ill-conditioning, so the Cholesky factorization gives a convenient way of doing this. We shall return to this topic in Chapter 9.

4.9 The Singular Value Decomposition

As we have already seen, various factorizations of a square matrix \mathbf{A} are important for computing a variety of quantities and for a theoretical understanding of properties of \mathbf{A} . It is of interest to know if similar decompositions exist for rectangular matrices, since the *design matrices* \mathbf{X} in regression analysis are generally not square. A factorization which is playing an increasingly important role in scientific calculations is the *singular value decomposition* (SVD), which may be viewed as a generalization of the spectral theorem for square matrices.

Theorem 4.18 (Singular value decomposition) *Let \mathbf{A} be a real $n \times m$ matrix. Then there exist two orthogonal matrices \mathbf{U}, \mathbf{V} such*

$$\mathbf{V}^T \mathbf{A} \mathbf{U} = \mathbf{F} \quad (4.200)$$

where \mathbf{F} is a diagonal rectangular $n \times m$ matrix of the form

$$\mathbf{F} = \left[\begin{array}{cccc|c} \mu_1 & 0 & \cdots & 0 & \\ 0 & \mu_2 & \cdots & 0 & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdots & \cdot & \mathbf{0} \\ \cdot & \cdot & & \cdot & \\ 0 & \cdot & \cdots & \mu_r & \\ \hline & & & \mathbf{0} & \mathbf{0} \end{array} \right] \quad (4.201)$$

where $F_{ii} = \mu_i, i = 1, 2, \dots, r$. The numbers $\mu_i, 1 \leq i \leq r$ are called the *singular values* of \mathbf{A} . They are real and positive and can be arranged so that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > 0$ where r is the rank of \mathbf{A} .

Proof. We give an outline of the proof, full details can be found in [112, 103]. Consider the square $m \times m$ matrix $\mathbf{A}^T \mathbf{A}$. It is symmetric and generally positive semidefinite. From the spectral theorem there is an orthogonal matrix \mathbf{U} such that

$$\mathbf{U}^T \mathbf{A}^T \mathbf{A} \mathbf{U} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r, 0, 0, \dots, 0) \quad (4.202)$$

where the elements of $\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r, 0, 0, \dots, 0)$ are the eigenvalues of $\mathbf{A}^T \mathbf{A}$. Hence, $\lambda_i > 0, 1 \leq i \leq r$, and can be arranged in decreasing order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$. Define $\mu_i = \sqrt{\lambda_i}, 1 \leq i \leq n$, and let

$$\mathbf{D} = \text{diag}(\mu_1, \mu_2, \dots, \mu_r, 0, 0, \dots, 0). \quad (4.203)$$

Hence (4.202) can be written as

$$\mathbf{U}^T \mathbf{A}^T \mathbf{A} \mathbf{U} = \mathbf{D}^2. \quad (4.204)$$

Letting $\mathbf{W} = \mathbf{A} \mathbf{U}$ (4.204) becomes

$$\mathbf{W}^T \mathbf{W} = \mathbf{D}^2 \quad (\mathbf{W} \text{ is } n \times m). \quad (4.205)$$

Letting $\mathbf{w}_j \in \mathbb{R}^n, 1 \leq j \leq m$, be the j -th column of \mathbf{W} it then follows from (4.205) that

$$\langle \mathbf{w}_j, \mathbf{w}_j \rangle = \begin{cases} \mu_i^2, & 1 \leq j \leq r \\ 0, & j > r \end{cases} \quad (4.206)$$

and $\mathbf{w}_j = \mathbf{0}$ for $j > r$. From (4.205) it follows that $\mathbf{w}_j, 1 \leq j \leq r$, are orthogonal, hence the first r columns of \mathbf{W} are linearly independent so that $r \leq n$.

Define

$$\mathbf{v}_j = \frac{\mathbf{w}_j}{\mu_j}, \quad 1 \leq j \leq r. \quad (4.207)$$

From (4.206) this is an orthonormal set in \mathbb{R}^n . Using the *Gram-Schmidt process* this can be completed to an orthonormal basis $\mathbf{v}_j, 1 \leq j \leq n$ for \mathbb{R}^n . Let

$$\mathbf{V} = [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_n] \quad (4.208)$$

and observe that $\mathbf{V} \mathbf{F} = \mathbf{W}$.

In fact,

$$\begin{aligned} \mathbf{V} \mathbf{F} &= [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_n] \left[\begin{array}{cccc|c} \mu_1 & 0 & \cdots & 0 & \\ 0 & \mu_2 & \cdots & 0 & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdots & \cdot & \\ \cdot & \cdot & & \cdot & \\ 0 & \cdot & \cdots & \mu_r & \\ \hline & 0 & & & 0 \end{array} \right] \\ &= \left[\begin{array}{cccc} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{array} \right] \left[\begin{array}{cccc|c} \mu_1 & 0 & \cdots & 0 & \\ 0 & \mu_2 & \cdots & 0 & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & \cdots & \cdot & \\ \cdot & \cdot & & \cdot & \\ 0 & \cdot & \cdots & \mu_r & \\ \hline & 0 & & & 0 \end{array} \right] \end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} v_{11}\mu_1 & v_{12}\mu_2 & \cdots & v_{1n}\mu_r \\ v_{21}\mu_1 & v_{22}\mu_2 & \cdots & v_{2n}\mu_r \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1}\mu_1 & v_{n2}\mu_2 & \cdots & v_{nn}\mu_r \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= [\mu_1 \mathbf{v}_1 | \mu_2 \mathbf{v}_2 | \cdots | \mu_r \mathbf{v}_r | 0 | 0 | \cdots | 0] \\
&= [\mathbf{w}_1 | \mathbf{w}_2 | \cdots | \mathbf{w}_r | 0 | 0 | \cdots | 0] = \mathbf{W}.
\end{aligned} \tag{4.209}$$

Thus,

$$\mathbf{V}\mathbf{F} = \mathbf{W} = \mathbf{A}\mathbf{U}. \tag{4.210}$$

Since \mathbf{V} is orthogonal, $\mathbf{V}^{-1} = \mathbf{V}^T$ so that $\mathbf{V}^T \mathbf{A}\mathbf{U} = \mathbf{F}$ as required. ■

Although we will return to this subject in Chapter 9 we make a few additional comments concerning the SVD. For the regression models considered in this text $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ where \mathbf{X} is a full rank $n \times m$ matrix with $n \geq m$ (usually $n > m$). In this case $r = m$ and \mathbf{F} is of the form

$$\mathbf{F} = \begin{bmatrix} \mu_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_m \\ \hline 0 & 0 & \cdots & 0 \end{bmatrix} \tag{4.211}$$

and the columns of \mathbf{W} are all nonzero. From $\mathbf{W} = \mathbf{A}\mathbf{U}$, these span the column space of \mathbf{A} . In fact, from the definition of \mathbf{W} ,

$$\mathbf{A}\mathbf{u}_j = \mu_j \mathbf{v}_j, \quad 1 \leq j \leq m \tag{4.212}$$

where \mathbf{u}_j is the j -th column of \mathbf{U} . The vectors $\mathbf{u}_j, 1 \leq j \leq m$, are called the *right singular vectors* of \mathbf{A} and $\mathbf{v}_j, 1 \leq j \leq m$, are the *left singular vectors* of \mathbf{A} . In fact, from (4.210)

$$\mathbf{U}^T \mathbf{A}^T \mathbf{A} \mathbf{U} = \mathbf{U}^{-1} \mathbf{A}^T \mathbf{A} \mathbf{U} = \mathbf{U}^{-1} \mathbf{A}^T \mathbf{W} = \mathbf{F}. \tag{4.213}$$

Thus, $\mathbf{A}^T \mathbf{W} = \mathbf{F}\mathbf{U}$ so that

$$\mathbf{A}^T \mathbf{v}_i = \mu_i \mathbf{u}_i, \quad 1 \leq i \leq m. \tag{4.214}$$

From (4.213) and (4.214), $\{\mathbf{u}_i\}_{i=1}^m$ and $\{\mathbf{v}_i\}_{i=1}^m$, play a role similar to that of the eigenvectors of a symmetric matrix.

When \mathbf{A} is an $m \times m$ invertible square matrix, then $\mu_i > 0, 1 \leq i \leq m$, and then the ratio

$$\kappa = \mu_1 / \mu_m \tag{4.215}$$

is called the *condition number* of \mathbf{A} . Its importance derives from the following fact. Suppose we wish to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$. In practice \mathbf{b} may be subject to various errors such as round-off errors. Thus rather than solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ we solve $\mathbf{A}\hat{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$ where $\Delta\mathbf{b}$

is the error in \mathbf{b} . The relative error in the computed solution $\|\mathbf{x} - \hat{\mathbf{x}}\| / \|\hat{\mathbf{x}}\|$ is bounded by $\kappa \|\Delta \mathbf{b}\| / \|\mathbf{b}\|$ giving [112]

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}}\|} \leq \frac{\kappa \|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}. \quad (4.216)$$

Thus κ measures the degree to which errors are magnified by the solution process. (From (4.215) $\kappa \geq 1$). If κ is large (this depends on the precision of one's computer arithmetic and the required number of digits) we say that \mathbf{A} is *ill-conditioned* otherwise it is *well-conditioned*. As we shall see in Chapter 9 the conditioning problem plays a major role in evaluating the validity of a regression model.

It is of interest to determine what types of matrices are well-conditioned and particularly those that are *optimally conditioned* with $\kappa = 1$. Since the singular values of \mathbf{A} are the eigenvalues of $\mathbf{A}^T \mathbf{A}$, if $\mathbf{A} = \mathbf{I}_m$, then $\mathbf{A}^T \mathbf{A} = \mathbf{I}_m^2 = \mathbf{I}_m$ and $\lambda_i(\mathbf{I}_m) = 1$, $1 \leq i \leq m$. Hence, $\kappa = 1$. More generally, any $m \times m$ diagonal matrix with constant diagonal elements has $\kappa = 1$.

If \mathbf{A} is orthogonal, then $\mathbf{A}^T \mathbf{A} = \mathbf{I}_m$ so again $\lambda_i(\mathbf{A}^T \mathbf{A}) = 1$, $1 \leq i \leq m$, and $\kappa = 1$. Hence, orthogonal matrices are optimally conditioned and this suggests that for numerical purposes one try to work with matrices that are orthogonal or close to it. The farther away a matrix is from being orthogonal - i.e., whose columns are not perpendicular vectors, it will be more poorly conditioned. The closer the angles between the columns are to zero, generally the larger the condition number. As we shall see, this geometric condition translates into the statistical notion of correlation.

4.9.1 The QR Decomposition

As indicated above, numerical calculations with orthogonal matrices are important so we now consider one further decomposition of \mathbf{A} in terms of an orthogonal matrix.

Theorem 4.19 (QR decomposition) *Let \mathbf{A} be an $n \times m$ matrix ($n \geq m$) of rank m , then \mathbf{A} can be factored as*

$$\mathbf{A} = \mathbf{Q}\mathbf{R} \quad (4.217)$$

where \mathbf{Q} is an $n \times m$ matrix with orthogonal columns and \mathbf{R} is an $m \times m$ upper triangular matrix.

The proof of (4.217) is a direct consequence of the Gram-Schmidt process. We will illustrate this for a 3×3 matrix, the proof in general follows along the same lines.

Suppose

$$\mathbf{A} = [\mathbf{a}_1 | \mathbf{a}_2 | \mathbf{a}_3] \quad (4.218)$$

where \mathbf{a}_i , $1 \leq i \leq 3$, are the columns of \mathbf{A} . By the Gram-Schmidt process there is a set of three orthogonal vectors \mathbf{w}_i , $1 \leq i \leq 3$, such that

$$\begin{aligned} \mathbf{w}_1 &= c_1 \mathbf{a}_1, \\ \mathbf{w}_2 &= c_2 \mathbf{a}_1 + c_3 \mathbf{a}_2, \\ \mathbf{w}_3 &= c_4 \mathbf{a}_1 + c_5 \mathbf{a}_2 + c_6 \mathbf{a}_3. \end{aligned} \quad (4.219)$$

Letting $\mathbf{w}_i = (w_{i1}, w_{i2}, w_{i3})^T$ and $\mathbf{a}_i = (a_{i1}, a_{i2}, a_{i3})^T$ writing out (4.219) explicitly shows that

$$\mathbf{Q} = [\mathbf{w}_1 | \mathbf{w}_2 | \mathbf{w}_3] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} c_1 & c_2 & c_4 \\ 0 & c_3 & c_5 \\ 0 & 0 & c_6 \end{bmatrix} = \mathbf{A}\mathbf{S} \quad (4.220)$$

where \mathbf{S} is upper triangular. Since \mathbf{A} and \mathbf{Q} are invertible by assumption, so is \mathbf{S} . Hence,

$$\mathbf{A} = \mathbf{Q}\mathbf{S}^{-1}. \quad (4.221)$$

Since the inverse of an upper triangular matrix is upper triangular, $\mathbf{S}^{-1} \equiv \mathbf{R}$ is upper triangular (4.217) follows [112]. ■

The importance of the \mathbf{QR} decomposition in regression analysis stems from the fact that \mathbf{R} can be produced numerically stably by applying a sequence of orthogonal transformations to \mathbf{A} and this can be used in calculating solutions to

$$\mathbf{X}^T \mathbf{X} \mathbf{x} = \mathbf{y} \quad (4.222)$$

by forming the \mathbf{QR} decomposition of \mathbf{X} . In fact, if $\mathbf{X} = \mathbf{QR}$ then $\mathbf{X}^T = \mathbf{R}^T \mathbf{Q}^T$ so

$$\mathbf{X}^T \mathbf{X} = \mathbf{R}^T \mathbf{Q}^T \mathbf{QR}. \quad (4.223)$$

Because the columns of \mathbf{Q} are orthogonal, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_m$ so that

$$\mathbf{X}^T \mathbf{X} = \mathbf{R}^T \mathbf{R} \quad (4.224)$$

By the triangularity of \mathbf{R} , we see that (4.224) is just the Cholesky decomposition of $\mathbf{X}^T \mathbf{X}$ with $\mathbf{L} = \mathbf{R}^T$ in (4.223). For numerical stability, this is generally the preferred method of obtaining the Cholesky factorization, hence the solution of (4.224). Further applications of these results will be given as we proceed.

4.10 Exercises

4.1 For $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 4 & -1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 0 & 4 & 2 \\ -1 & -1 & 0 & 1 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ -1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$,

find

(a) \mathbf{AB} (b) $\mathbf{A}^T \mathbf{B}$ (c) \mathbf{Ay} (d) \mathbf{Bx} (e) $\mathbf{y}^T \mathbf{A}^T \mathbf{Ay}$ (f) $\mathbf{A}^2 - 2\mathbf{A} + \mathbf{I}^2$

4.2 Let $\mathbf{X} = \begin{bmatrix} 1 & 2 & 4 \\ 0 & -1 & -2 \\ -1 & 0 & 3 \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} 2 & 0 & 0 \\ -2 & 5 & 0 \\ 0 & -3 & 1 \end{bmatrix}$.

(a) Find \mathbf{X}^2 , \mathbf{Y}^2 , \mathbf{XY} , \mathbf{YX} .

(b) Show that $(\mathbf{X} + \mathbf{Y})^2 = \mathbf{X}^2 + \mathbf{XY} + \mathbf{YX} + \mathbf{Y}^2$.

4.3 Verify that

(a) If $\mathbf{A} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, then $\mathbf{A}^2 = \mathbf{A}$.

(b) If $\mathbf{B} = \begin{bmatrix} 6 & -4 \\ 9 & -6 \end{bmatrix}$, then \mathbf{B}^2 is $\mathbf{0}$.

4.4 For $\mathbf{A} = \begin{bmatrix} 1 & 8 & 2 \\ 6 & 9 & -7 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 6 \\ 2 & 8 \\ 3 & 4 \end{bmatrix}$.

Find $\mathbf{A} + \mathbf{B}^T$ and $\mathbf{A}^T + \mathbf{B}$, and explain the relationship between two sums.

4.5 Prove the properties (i)-(vi) of the trace of a matrix.

4.6 Prove that if $\mathbf{A} = \text{diag}(a_i)$, $1 \leq i \leq n$, and $a_i \neq 0$, $1 \leq i \leq n$, then $\mathbf{A}^{-1} = \text{diag}(1/a_i)$, $1 \leq i \leq n$. In particular, $\mathbf{I}_n^{-1} = \mathbf{I}_n$.

4.7 Prove Theorem 4.6 (i)-(v).

4.8 Let $\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 4 & 1 \\ 3 & 0 & 2 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 3 & 4 & 0 \\ 0 & -1 & 2 \\ 1 & 0 & 1 \end{bmatrix}$.

(a) Partition \mathbf{A} and \mathbf{B} as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix},$$

where both \mathbf{A}_{11} and \mathbf{B}_{11} have the same dimension 2×2 .

(b) Find \mathbf{AB} both with and without the partitioning, to verify the validity of multiplication of partitioned matrices.

(c) Find \mathbf{AB}^T by showing that

$$\mathbf{B}^T = \begin{bmatrix} \mathbf{B}_{11}^T & \mathbf{B}_{12}^T \\ \mathbf{B}_{21}^T & \mathbf{B}_{22}^T \end{bmatrix}.$$

4.9 (a) Consider the vectors $\mathbf{b} = (2, -1, 4, 0)^T$ and $\mathbf{d} = (-1, 3, -2, 1)^T$. Show that

$$(\mathbf{b}^T \mathbf{d})^2 \leq (\mathbf{b}^T \mathbf{b}) (\mathbf{d}^T \mathbf{d}).$$

(b) Consider the vectors $\mathbf{b} = (-3, 4)^T$ and $\mathbf{d} = (2, 1)^T$. Show that

$$(\mathbf{b}^T \mathbf{d})^2 \leq (\mathbf{b}^T \mathbf{\Lambda} \mathbf{b}) (\mathbf{d}^T \mathbf{\Lambda}^{-1} \mathbf{d})$$

where

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}.$$

4.10 Find the inner product of $\mathbf{y} - (\mathbf{x}^T \mathbf{y} / \mathbf{x}^T \mathbf{x}) \mathbf{x}$ with itself.

4.11 Which of the following transformation $\mathbf{A} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ are linear? Let $\mathbf{A}(x, y, z)$ be given

- (a) $(x, y + 1)$ (b) (x, x) (c) $(0, 0)$
 (d) $(xy, 0)$ (e) $(y, x + \sqrt{2}y)$ (f) $(y, x + \sqrt{y})$

4.12 Let $\mathbf{T} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be a linear transformation.

(a) If $\mathbf{T}(2, 3) = (-1, 2)$ and $\mathbf{T}(1, 2) = (5, 2)$, what is $\mathbf{T}(4, 7)$? [Hint: Write $(4, 7)$ as a linear combination of $(2, 3)$ and $(1, 2)$.]

(b) If $\mathbf{T}(\mathbf{v}_1) = \mathbf{v}_1$ and $\mathbf{T}(\mathbf{w}_1) = \mathbf{w}_1$, what is $\mathbf{T}(2\mathbf{v} - 3\mathbf{w})$?

(c) Suppose $\mathbf{T}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and $\mathbf{T}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$. Find a matrix \mathbf{A} such that \mathbf{T} is the action of \mathbf{A} on \mathbb{R}^2 .

4.13 In a genetic study, the *generation matrix*, $\mathbf{A} = \begin{bmatrix} 1 & 1/2 \\ 0 & 1/2 \end{bmatrix}$ is given by Kempthorne (1957, p. 120) [68]. Given that $\mathbf{f}^{(i)} = \mathbf{A}\mathbf{f}^{(i-1)}$ ($i \geq 1$) show that

$$\mathbf{f}^{(2)} = \begin{bmatrix} 1 & 3/4 \\ 0 & 1/4 \end{bmatrix} \mathbf{f}^{(0)}, \quad \mathbf{f}^{(3)} = \begin{bmatrix} 1 & 7/8 \\ 0 & 1/8 \end{bmatrix} \mathbf{f}^{(0)}$$

and

$$\mathbf{f}^{(n)} = \mathbf{A}^n \mathbf{f}^{(0)} = \begin{bmatrix} 1 & 1 - 2^{-n} \\ 0 & 2^{-n} \end{bmatrix} \mathbf{f}^{(0)},$$

where $\mathbf{f}^{(0)}$ is an arbitrary vector in \mathbb{R}^2 .

4.14 Let the row vector $\mathbf{1}_4^T = (1, 1, 1, 1)$, $\mathbf{1}_3^T = (1, 1, 1)$ and $\mathbf{x}^T = (3, 6, 8, -2)$. Verify

(a) $\mathbf{1}_4^T \mathbf{x} = \mathbf{x}^T \mathbf{1}_4$

(b) $\mathbf{1}_4 \mathbf{1}_3 = \mathbf{J}_{4 \times 3}$, having all elements unity.

4.15 A square matrix \mathbf{J}_n which is quite useful and a variant thereof is given by:

$$\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n^T \quad \text{with} \quad \mathbf{J}_n^2 = n \mathbf{J}_n;$$

and

$$\bar{\mathbf{J}}_n = \frac{1}{n} \mathbf{J}_n \quad \text{with} \quad \bar{\mathbf{J}}_n^2 = \bar{\mathbf{J}}_n.$$

Particularly, a *centering matrix*, \mathbf{C}_n is defined by

$$\mathbf{C}_n = \mathbf{I} - \bar{\mathbf{J}}_n = \mathbf{I} - \frac{1}{n} \mathbf{J}_n.$$

Verify that $\mathbf{C}_n = \mathbf{C}_n^T = \mathbf{C}_n^2$, $\mathbf{C}_n \mathbf{1}_n = \mathbf{0}$ and $\mathbf{C}_n \mathbf{J}_n = \mathbf{J}_n \mathbf{C}_n = \mathbf{0}$.

4.16 Let \mathbf{A} be a symmetric matrix.

Prove that if $\mathbf{A} = \mathbf{A}^{-1}$, then $\det(\mathbf{A}) = \pm 1$.

4.17 Let $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$. Show that

(a) $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.

(b) $\det(\mathbf{A}^{-1}) = [\det(\mathbf{A})]^{-1}$.

4.18 Let $\mathbf{S} = \begin{bmatrix} a & b & b \\ b & a & b \\ b & b & a \end{bmatrix}$, where $b > 0$.

- (a) Find the value of a such that $|\mathbf{S}| = 0$.
 (b) Find \mathbf{S}^{-1} .

4.19 Let $\mathbf{A} = \begin{bmatrix} -2 & 3 & -1 \\ 1 & 2 & -1 \\ -2 & -1 & 1 \end{bmatrix}$.

- (a) Calculate the transpose of \mathbf{A}^{-1} and inverse of \mathbf{A}^T .
 (b) Calculate the inverse of \mathbf{A}^{-1} .

4.20 Let \mathbf{A} be an $n \times n$ matrix such that $\mathbf{A}^T = \mathbf{A}^{-1}$.

- (a) Show that any matrix of the form $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$, where $a^2 + b^2 = 1$, has this property.
 (b) Show that $\det(\mathbf{A}) = \pm 1$ for such a matrix.

4.21 Prove the followings.

- (a) If $\mathbf{AB} = \mathbf{I}_n$, then $\det(\mathbf{A}) \neq 0$ and $\det(\mathbf{B}) \neq 0$.
 (b) For \mathbf{A} and \mathbf{B} symmetric, $[(\mathbf{AB})^T]^{-1} = \mathbf{A}^{-1}\mathbf{B}^{-1}$.
 (c) Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, then \mathbf{P} is symmetric and idempotent.

4.22 Find the characteristic polynomial, the eigenvalues, and their associated eigenvectors for each of the following matrices.

(a) $\begin{bmatrix} 2 & 1 \\ -1 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & -2 \\ 0 & -2 & 0 \\ -2 & 0 & 3 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ 3 & 2 & -2 \end{bmatrix}$.

4.23 Let \mathbf{A} be an upper triangular matrix. Prove that all diagonal entries of \mathbf{A} are eigenvalues. (It is also called a *proper value*). [Hint: The determinant of an upper triangular matrix is the product of its diagonal elements.]

4.24 Prove that λ is an eigenvalue of $\mathbf{A}_{n \times n}$ if and only if $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, where \mathbf{I} is the $n \times n$ identity matrix.

4.25 Define an observation vector (or data vector) $\mathbf{x}^T = (x_1, x_2, \dots, x_n)$.

- (a) Show that the mean vector $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{x}^T\mathbf{1} = \frac{1}{n}\mathbf{1}^T\mathbf{x}$.
 (b) Find the deviation vector \mathbf{C} such that $\mathbf{C}\mathbf{x} = \mathbf{x}^T - \bar{\mathbf{x}}\mathbf{1}^T$.
 (c) Show that $\sum_{i=1}^n (x_i - \bar{x})^2 = \mathbf{x}^T\mathbf{x} - n\bar{x}^2 = \mathbf{x}^T\mathbf{C}\mathbf{x}$.

4.26 Let X_1, X_2, \dots, X_n be a random sample with mean θ , and variance σ^2 . Find the expected value of the quadratic form

$$Q = (X_1 - X_2)^2 + (X_2 - X_3)^2 + \dots + (X_{n-1} - X_n)^2.$$

4.27 Find matrices \mathbf{A} such that the following quadratic forms are given by the product $\mathbf{x}^T \mathbf{A} \mathbf{x}$.

(a) $2x^2 - 6xy - y^2$.

(b) $4x_1^2 + 3x_1x_2 - 2x_2^2$.

(c) $x^2 - \frac{1}{2}y^2 + 4z^2$.

(d) $2x_1^2 - 2x_1x_2 + x_2^2 + 4x_1x_3 - 3x_3^2$.

4.28 Let $\mathbf{A} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$ be a 2×2 symmetric matrix. Prove that \mathbf{A} is positive definite if and only if $\det(\mathbf{A}) > 0$ and $a > 0$.

4.29 Classify the following matrices as positive definite or positive semidefinite:

(a) $\begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & -1 \\ 2 & -1 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix}$.

4.30 Which of the following matrices are diagonalizable?

(a) $\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 4 \\ 1 & -2 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 & -2 \\ 4 & 0 & 4 \\ 1 & -1 & 4 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & 2 \\ 0 & 0 & 2 \end{bmatrix}$.

4.31 Let \mathbf{A} be a symmetric matrix such that $\mathbf{x}^T \mathbf{A} \mathbf{x}$ yields the given quadratic form $x^2 + 4xy + y^2$.

(a) Find the eigenvalues and eigenvectors of \mathbf{A} .

(b) Find an orthogonal matrix \mathbf{P} such that $\mathbf{P}^T \mathbf{A} \mathbf{P}$ is a diagonal matrix.

4.32 Let $\mathbf{x} = (2, 1, -3)$, $\mathbf{y} = (-5, 1, 2)$, and $\mathbf{z} = (1, -4, 2)$. If possible, compute

(a) \mathbf{xy}^T (b) $\mathbf{z}^T \mathbf{z}$ (c) $(\mathbf{yx}^T) \mathbf{z}$ (d) $3\mathbf{y}^T$ (e) $\mathbf{yx}^T \mathbf{yx}^T$

4.33 Consider the following linear system:

$$\begin{cases} 2x_1 & +3x_2 & -3x_3 & +x_4 & +x_5 & = 7 \\ 3x_1 & & +2x_3 & & +3x_5 & = -2 \\ 2x_1 & +3x_2 & & -4x_4 & & = 3 \\ & & x_3 & +x_4 & +x_5 & = 5. \end{cases}$$

(a) Find the coefficient matrix.

(b) Write the linear system in matrix form.

(c) Find the augmented matrix.

4.34 Using Gaussian elimination method, solve the following sets of simultaneous equations.

(a) $\begin{cases} x - y = 2 \\ 2x + y = 1. \end{cases}$ (b) $\begin{cases} x_1 + x_2 + 2x_3 = -1 \\ x_1 - 2x_2 + x_3 = -5 \\ 3x_1 + x_2 + x_3 = 3. \end{cases}$ (c) $\begin{cases} -x_1 + 2x_2 - x_3 = -2 \\ 4x_1 - 2x_2 + 2x_3 = 2 \\ 3x_1 - 4x_3 = -1. \end{cases}$

4.35 Write each of the following as a linear system in matrix form.

$$(a) \ x_1 \begin{pmatrix} -1 \\ 2 \end{pmatrix} + x_2 \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

$$(b) \ x_1 \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ -3 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix}.$$

4.36 Let $\mathbf{x}_1 = (2, -1, 1)^T$, $\mathbf{x}_2 = (4, -7, -1)^T$, $\mathbf{x}_3 = (1, 2, 2)^T$ belong to the solution space of $\mathbf{A}\mathbf{x} = \mathbf{0}$ where \mathbf{A} is a nonzero 3×3 matrix. Is $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ linearly independent?

4.37 Decide whether or not the following vectors are linearly independent, by solving $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$;

$$(a) \ \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \ \mathbf{v}_2 = \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix}, \ \mathbf{v}_3 = \begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix}.$$

$$(b) \ \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}, \ \mathbf{v}_2 = \begin{pmatrix} 2 \\ -2 \\ 4 \end{pmatrix}, \ \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 4 \end{pmatrix}.$$

4.38 Let

$$\mathbf{t}^T = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \ \mathbf{u}^T = (1, 0, -1, 0), \ \mathbf{v}^T = \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}, 0\right)$$

(a) Identify \mathbf{t}^T , \mathbf{u}^T and \mathbf{v}^T are orthogonal or orthonormal.

(b) Find the angles of each pair of vector.

(c) Compute the length of each vector.

4.39 Let $\mathbf{x}^T = [1, -3, 2]$, $\mathbf{y}^T = [2, -4, 5]$.

(a) Find the vector $\hat{\mathbf{y}}$ such that $\hat{\mathbf{y}} = b\mathbf{x}$ where $b = \langle \mathbf{x}, \mathbf{y} \rangle / \|\mathbf{x}\|^2$.

(b) Find $\hat{\mathbf{y}} - \mathbf{y}$ and show that $(\hat{\mathbf{y}} - \mathbf{y}) \perp \mathbf{x}$.

4.40 Draw a picture and prove the parallelogram law:

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2).$$

4.41 Let $\mathbf{x}_1 = [1, 1, 0]^T$, $\mathbf{x}_2 = [1, 0, 1]^T$ and $\mathbf{x}_3 = [0, 1, 1]^T$. Find the orthonormal vectors \mathbf{q}_1 , \mathbf{q}_2 and \mathbf{q}_3 .

4.42 Find an orthonormal basis for the subspace of \mathbb{R}^3 consisting of all vectors of the form

(a) $(a, a + b, b)$ (b) (a, b, c) such that $a + b + c = 0$.

4.43 Consider the Euclidean space \mathbb{R}^4 and let \mathcal{W} be the subspace that has $\mathcal{S} = \{[1, 1, -1, 0], [0, 2, 0, 1]\}$ as a basis. Use the Gram-Schmidt process to obtain an orthonormal basis for \mathcal{W} .

4.44 Find the factorization $\mathbf{A} = \mathbf{LDL}^T$ where \mathbf{D} is a diagonal matrix, and then the two Cholesky factors in

$$\left(\mathbf{LD}^{1/2}\right)\left(\mathbf{LD}^{1/2}\right)^T, \text{ for}$$

$$\mathbf{A} = \begin{bmatrix} 2 & 6 \\ 6 & 21 \end{bmatrix}.$$

4.45 Find the QR decomposition of

$$(a) \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix} \quad (b) \begin{bmatrix} 2 & -1 \\ -1 & 3 \\ 0 & 1 \end{bmatrix} \quad (c) \begin{bmatrix} 1 & 2 \\ -1 & -2 \\ 1 & 1 \end{bmatrix} \quad (d) \begin{bmatrix} -1 & 2 & 0 \\ 1 & 0 & 2 \\ -1 & -2 & 2 \end{bmatrix}.$$

4.46 Use the Gram-Schmidt process to

$$\mathbf{u} = (0, 0, 1)^T, \mathbf{v} = (0, 1, 1)^T, \mathbf{w} = (1, 1, 1)^T$$

and write the result in the form $\mathbf{A} = \mathbf{QR}$.

4.47 With the same matrix \mathbf{A} , and with $\mathbf{b} = (1, 1, 1)^T$, use $\mathbf{A} = \mathbf{QR}$ to solve the least squares problem $\mathbf{Ax} = \mathbf{b}$.

4.48 Find the singular value decomposition (SVD) of the matrices (b) and (c) in Exercise 4.45.

4.49 Let $\mathbf{X}_{3 \times 1}$ be $\mathbf{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 4 \end{bmatrix}.$$

- (a) Find the marginal distribution of X_1 .
- (b) Find the joint distribution of X_2 and X_3 .
- (c) Find the distribution of $Z = 2X_1 - X_2 + 3X_3$.
- (d) Find the conditional distribution of X_1 , given that $X_2 = x_2$.
- (e) Find the conditional distribution of (X_1, X_2) given that $X_3 = x_3$.
- (f) Are X_1 and X_3 independent? Why?
- (g) Find a 2×1 vector \mathbf{a} such that X_2 and $X_2 - \mathbf{a}^T \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ are independent.
- (h) Find the covariance matrix of (X_1, X_2) given that $X_3 = x_3$. Using this find the correlation of (X_1, X_2) given that $X_3 = x_3$, $\rho_{1,2|3}$.

4.50 Using $\boldsymbol{\Sigma}$ given in Ex. 4.49, find the covariance of Z_1 and Z_2 , where

$$\begin{aligned} Z_1 &= X_1 - 2X_2 + 3X_3 - 6 \\ Z_2 &= 2X_1 + 3. \end{aligned}$$

4.51 Suppose $\mathbf{x}_{p \times 1} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Find the moment generating function of \mathbf{y} , where $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$.

4.52 Suppose $\mathbf{x}_{p \times 1} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite, the pdf is given by

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}Q\right)$$

where $Q = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$.

(a) Show that the mean vector \mathbf{x} is the solution for \mathbf{x} to $\partial f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) / \partial \mathbf{x} = \mathbf{0}$.

(b) Find the solution for \mathbf{x} to $\partial Q / \partial \mathbf{x} = \mathbf{0}$. Is this equivalent to the one you obtained in (a)?

4.53 Suppose a random vector $\mathbf{x}_{2 \times 1}$ is distributed $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the pdf is given by

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-1} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}Q\right)$$

where $Q = 3x_1^2 + 2x_2^2 - 4x_1x_2 + 14x_1 - 8x_2 + 10$. (Note: This is called the bivariate normal distribution).

(a) Find $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$.

(b) Find $E(X_1 | X_2 = x_1)$.

Chapter 5

Multiple Regression

5.1 Introduction

In this chapter we extend the ideas developed in Chapter 3 to fit models of data when there is more than one independent variable. As we shall see, the use of matrix algebra greatly simplifies many of the calculations and will be used consistently throughout the chapter.

5.2 The General Linear Model

In Chapter 3 we studied the problem of using a single independent variable x to aid in estimating the mean value $E(Y_x)$ of a family of random variables Y_x . As we pointed out in Example 3.21, a single variable may be inadequate to explain the variation in some variable Y . In this chapter we consider the problem of regressing Y on one or more variables as a generalization of the techniques developed in Chapter 3. This topic is usually called *multiple regression* and the resulting models *multiple regression models* or *general linear models*.

As a specific situation, we might expect that the income a person earns depends on a number of factors such as the number of years of education, the type of job, age, sex (kind; not how much) and no doubt others. Likewise, the price of a house generally depends on its size, location, configuration and number of rooms, age, lot size and others. Many other examples of such phenomena will be given as we proceed.

Suppose now that we have $m \geq 1$ independent variables x_1, x_2, \dots, x_m , denoted in vector form as $(x_1, x_2, \dots, x_m) \equiv \mathbf{x}$. Let $Y_{\mathbf{x}}$ be a random variable whose mean $E(Y_{\mathbf{x}}) = \mu_{\mathbf{x}}$ depends in some fashion on \mathbf{x} . Generalizing (3.1) we will assume for now that

$$Y_{\mathbf{x}} = \beta_0 + \sum_{j=1}^m \beta_j x_j + \varepsilon_{\mathbf{x}} \quad (5.1)$$

where $\beta_j, j = 0, 1, \dots, m$, are fixed, but generally unknown, real numbers, and $\varepsilon_{\mathbf{x}}$ is an error random variable satisfying $E(\varepsilon_{\mathbf{x}}) = 0$. Thus,

$$\mu_{\mathbf{x}} = E(Y_{\mathbf{x}}) = \beta_0 + \sum_{j=1}^m \beta_j x_j, \quad (5.2)$$

so that we are assuming that the dependence of $\mu_{\mathbf{x}}$ on the *regression coefficients* $(\beta_0, \beta_1, \dots, \beta_m)$ is linear. The model given by Eq. (5.1) is usually referred to as the *general linear model*.

As for the simple linear regression model, we refer to $\beta_j, 0 \leq j \leq m$, as the *regression coefficients*; the x_i 's as the *independent variables* (variables for short) or *regressors*. β_0 is usually called the *intercept* and $Y_{\mathbf{x}}$ the *dependent* or *observed (response) variable*. When a particular variable x_i is quantitative, that is, it assumes values in an interval of the real line, then β_i is sometimes called the *i-th slope*, since it represents the rate of change of $Y_{\mathbf{x}}$ in the direction of x_i . That is, β_i is the amount that $Y_{\mathbf{x}}$ changes when x_i increases by one unit when all the other variables are held fixed. (In calculus terms $\beta_i = \partial\mu_{\mathbf{x}}/\partial x_i$, the partial derivative of $\mu_{\mathbf{x}}$ with respect to x_i .) When x_i is qualitative, that is, it can assume only a discrete set of values, often only 0 and 1, then no such calculus interpretation of β_i is usually possible. Some authors refer to $\beta_j, 0 \leq j \leq m$, as *partial regression coefficients*, but we will not use this terminology here.

For the remainder of this chapter we shall assume that the x_i 's are deterministic, and can be measured without error. Similarly, it is assumed that the values of $Y_{\mathbf{x}}$ can be measured without error as well.

To illustrate the scope of the general linear model in practice we now present a number of general and specific models that have appeared in the literature. The reader will see that the general linear model encompasses a wide variety of statistical possibilities.

Example 5.1 In Example 3.4 we considered using a simple linear regression model to explain the delivery time for drinks to a customer as a function of the number of cases delivered. A further examination of the data suggested that the distance the delivery man had to walk could be of some significance. To investigate this possibility a linear model

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{\mathbf{x}} \quad (5.3)$$

was considered. Here, as in Example 3.4 $Y_{\mathbf{x}}$ is the delivery time in minutes, x_1 is the number of cases delivered and x_2 is the distance walked, in feet.

Example 5.2 The amount of water used in a production plant is quite large. In order to understand the factors determining the amount of monthly water usage, the cost control engineer considered using a linear regression model of the form

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_{\mathbf{x}} \quad (5.4)$$

where

- $Y_{\mathbf{x}}$ = monthly water usage in gallons,
- x_1 = average monthly temperature,
- x_2 = amount of production (pounds),
- x_3 = number of operating days in a month,
- x_4 = number of persons on the plant payroll.

Example 5.3 In exercise physiology an objective measure of aerobic fitness is the oxygen consumption in volume per unit body weight per unit time (Y). To determine if it was possible to predict this quantity, an experiment was conducted and a linear regression model

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon_{\mathbf{x}} \quad (5.5)$$

was proposed. Here the variables

- x_1 = age in years,
- x_2 = weight in kilograms,
- x_3 = time to run 1.5 miles,
- x_4 = resting pulse rate,
- x_5 = pulse rate at the end of the run,
- x_6 = maximum pulse rate during the run

were used to attempt to explain the results of the experiment.

Example 5.4 (Polynomial regression models) In the case of the simple linear regression model we pictured the data (x_i, y_i) as being roughly scattered about a straight line. In this case the model is linear in both the parameters (β_0, β_1) and the independent variable x . In the general linear model, the *linearity property* refers only to the linear behavior in the coefficients $(\beta_0, \beta_1, \dots, \beta_m)^T$. The model need not be linear in the independent variables.

To clarify this, suppose $x_i = x^i$ where x is some explanatory variable. Then consider the model

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m + \varepsilon_{\mathbf{x}}. \quad (5.6)$$

In this case, the model has a polynomial behavior in x , but is still linear in $(\beta_0, \beta_1, \dots, \beta_m)^T$. Such a model is called a *polynomial model of degree m* .

One can also use the general linear model, to model polynomial behavior in two or more variables. For example, a two variable *quadratic model* would be of the form

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon_{\mathbf{x}}. \quad (5.7)$$

We will discuss these models in greater detail in Chapter 7.

One of the most important aspects of the general linear model is its ability to describe both quantitative and qualitative behavior simultaneously.

Example 5.5 Suppose that it is felt that the salary that a person earns depends in a linear fashion on the number of years worked but that the rate of growth depends on a person's sex. (Such models can be useful in studying employment discrimination.) To test this possibility data were gathered and it was planned to fit a model. How should one proceed?

Of course one possibility is to fit two straight lines, one for each sex, and be done. But in a study of this type one of the things that we might be interested in is the ability to determine whether sex does have an influence on the rate of change of salary. Thus we will want to determine whether the salary models for males and females differ in their slopes. Although there are a number of ways of doing this, it is perhaps most convenient (and efficient) to have one model which will fit the male and female salary data simultaneously. This can be done using the general linear model by introducing a *dummy variable* x_2 where

$$x_2 = \begin{cases} 0, & \text{if person is female,} \\ 1, & \text{if person is male.} \end{cases} \quad (5.8)$$

Letting $x_1 =$ years worked, the model

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{\mathbf{x}}. \quad (5.9)$$

can be shown to model male and female salaries simultaneously. To see this, let $x_2 = 0$, then

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \varepsilon_{\mathbf{x}}$$

which represents the model for female salaries. If $x_2 = 1$, then

$$\begin{aligned} Y_{\mathbf{x}} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{\mathbf{x}} \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon_{\mathbf{x}} \\ &= \beta'_0 + \beta'_1 x_1 + \varepsilon_{\mathbf{x}} \end{aligned} \quad (5.10)$$

which represents the model for male salaries. Note that in this representation β_2 represents the difference between male and female starting salaries and β_3 represents the difference in annual rate increases. In this form it is easy to test for significant differences in these quantities. We shall take up such models in greater detail in Chapter 7. In the statistics literature these are sometimes called *analysis of covariance models* [88]. As we also shall see in Chapter 7, it is possible to consider models having only dummy variables, called *analysis of variance models*. These models historically have been treated as separate statistical models with their own terminology, methodology etc. [27]. However, increasingly one finds that such models are being analyzed as particular examples of the GLM [27, 88], which appears, at least to these writers to remove a great deal of mystery and complication from the subject.

Although the linearity assumption in (5.1) is theoretically restrictive, the proper choice of $x_i, 1 \leq i \leq m$ and $Y_{\mathbf{x}}$ often allows one to represent many physical phenomena quite adequately over appropriate ranges of the independent variables. Again we will deal with this form of the model until Chapter 7.

5.3 Least Squares Estimation

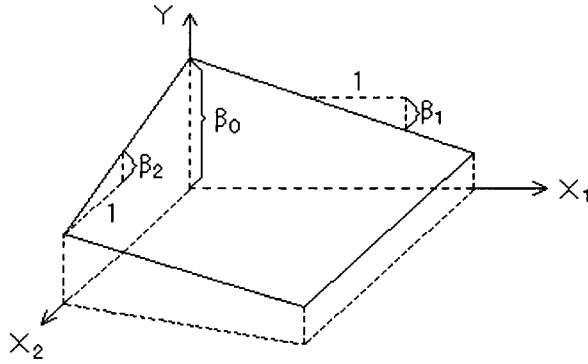
5.3.1 Estimating β

As for the case of simple linear regression, the first problem we deal with is that of estimating the regression coefficients $(\beta_0, \beta_1, \dots, \beta_m)$ and the residual variability of $\varepsilon_{\mathbf{x}}$. To do this we assume that n , not necessarily distinct, measurements have been taken of the independent variables $\mathbf{x} = (x_1, x_2, \dots, x_m)$. Let x_{ij} be value of the j -th variable for the i -th measurement. Then, letting $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$,

$$Y_{\mathbf{x}_i} \equiv Y_i = \beta_0 + \sum_{j=1}^m \beta_j x_{ij} + \varepsilon_i, \quad 1 \leq i \leq n, \quad (5.11)$$

where $\varepsilon_i \equiv \varepsilon_{\mathbf{x}_i}$.

The observed values of Y_i at \mathbf{x}_i will be denoted as usual by the lower case value y_i . Our basic problem is the estimation of $\beta = (\beta_0, \beta_1, \dots, \beta_m)^T$ from the pairs of data $(\mathbf{x}_i, y_i), 1 \leq i \leq n$. For the remainder of this chapter we shall assume that $n > m + 1$ so that we have at least one more measurement than the number of parameters $\{\beta_i\}_{i=0}^m$.

Figure 5.1: A Linear Regression Surface with $m = 2$

Geometrically, if we think of representing (\mathbf{x}_i, y_i) as a point in \mathbb{R}^{m+1} , then in analogy to simple linear regression, we may think of the estimation problem as one of fitting an $m + 1$ dimensional hyperplane to this scatter of points. An example of this situation for $m = 2$ is shown in Figure 5.1.

Before proceeding with the actual estimation details it is convenient to rewrite the model (5.1) in vector-matrix form. This notation, coupled with the techniques of linear algebra surveyed in Chapter 4, provides a powerful method for developing the properties of the general linear model.

As in Chapter 4 (and so far in this chapter) vectors and matrices will be denoted by boldface letters, and if necessary their dimensions will be denoted as in Eq. (5.16).

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ denote the vector of random observations and their actual values by

$$\mathbf{y} = (y_1, y_2, \dots, y_n)^T. \quad (5.12)$$

The error vector $\boldsymbol{\varepsilon}$ is given by

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T \quad (5.13)$$

and the vector of regression coefficients is denoted by

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)^T. \quad (5.14)$$

Let

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}_{n \times (m+1)} \quad (5.15)$$

where \mathbf{X} is called the *design matrix*. (This terminology is used even if the values of \mathbf{x}_i do not result from a pre-planned experiment [104, 45].) With these definitions (5.11) takes the form

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (m+1)} \boldsymbol{\beta}_{(m+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}. \quad (5.16)$$

For now we make the same assumptions on ε as these given in Section 3.2. That is, $\varepsilon_i, 1 \leq i \leq n$, are assumed to be independent, and each ε_i is $N(0, \sigma^2)$. Thus, the vector \mathbf{Y} has a multivariate normal $\mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ distribution. These assumptions, as in Chapter 3, lead naturally to choosing maximum likelihood estimation as a way of estimating $\boldsymbol{\beta}$, and then to least squares estimation for more general error models.

When the errors are independent $N(0, \sigma^2)$, the likelihood function of \mathbf{y} is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^n \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2\sigma^2} \left(y_i - \beta_0 - \sum_{j=1}^m x_{ij}\beta_j \right)^2 \right] \right\}. \quad (5.17)$$

Writing $\mu_i = \beta_0 + \sum_{j=1}^m x_{ij}\beta_j$, (5.17) takes the form

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 \right] \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left[-\frac{1}{2\sigma^2} \langle \mathbf{y} - \boldsymbol{\mu}, \mathbf{y} - \boldsymbol{\mu} \rangle \right] \end{aligned} \quad (5.18)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$ and $\langle \mathbf{y} - \boldsymbol{\mu}, \mathbf{y} - \boldsymbol{\mu} \rangle$ is the dot product of the vector $\mathbf{y} - \boldsymbol{\mu}$ with itself. Since $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, (5.18) becomes

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left[-\frac{1}{2\sigma^2} \langle \mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \rangle \right]. \quad (5.19)$$

From (5.19) we see that maximizing $f_{\mathbf{Y}}(\mathbf{y})$ with respect to $\boldsymbol{\beta}$ is the same as minimizing the sum of squares of the residuals

$$g(\boldsymbol{\beta}) = \langle \mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \rangle. \quad (5.20)$$

The usual approach to doing this minimization is to differentiate $g(\boldsymbol{\beta})$ with respect to $\beta_i, 0 \leq i \leq m$, and solve the resulting equations obtained by setting these derivatives to zero. This is the approach taken in Chapter 3 for the simple linear regression model (but, see the derivation in Section 3.2). However, one is still required to verify that the estimates so obtained actually do provide the minimum and this requires additional calculus. Due to the quadratic behavior of $g(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$, a purely algebraic approach to the minimization is possible. This is the route we follow. The calculus approach will be outlined in the Exercises.

Theorem 5.1 *Consider the linear model given by Equation (5.11). If the errors are independent $N(0, \sigma^2)$ random variables then $\hat{\boldsymbol{\beta}}$ is a maximum likelihood estimate of $\boldsymbol{\beta}$ if and only if $\hat{\boldsymbol{\beta}}$ is a solution to the set of linear equations*

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}. \quad (5.21)$$

If $\mathbf{X}^T \mathbf{X}$ is nonsingular, then (5.21) has a unique solution which is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (5.22a)$$

In this case the estimator $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (5.22b)$$

(Note that we use the same notation for the estimator and its values $\hat{\beta}$. This is standard practice, and should, we hope, cause no confusion.)

Note: In many problems, particularly those of analysis of variance type, (5.22b) can lead to *singular matrices* $\mathbf{X}^T \mathbf{X}$. Thus the MLE of β need not be unique, so that the assumption of nonsingularity of $\mathbf{X}^T \mathbf{X}$ is not superfluous.

We will show that the proof of Theorem 5.1 is a straightforward consequence of the identity

$$g(\beta) - g(\hat{\beta}) = \langle \mathbf{X}(\beta - \hat{\beta}), \mathbf{X}(\beta - \hat{\beta}) \rangle, \quad (5.23)$$

where $\hat{\beta}$ satisfies (5.22b) and the fact that (5.21) always has at least one solution.

To establish (5.23) consider

$$\begin{aligned} g(\beta) &= \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle \\ &= \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{y}, \mathbf{X}\beta \rangle + \langle \mathbf{X}\beta, \mathbf{X}\beta \rangle \\ &= \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{X}^T \mathbf{y}, \beta \rangle + \langle \mathbf{X}\beta, \mathbf{X}\beta \rangle. \end{aligned} \quad (5.24)$$

Using $\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \hat{\beta}$ in the middle term of (5.24), gives

$$\begin{aligned} g(\beta) &= \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{X}^T \mathbf{y}, \beta \rangle + \langle \mathbf{X}\beta, \mathbf{X}\beta \rangle \\ &= \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{X}\hat{\beta}, \mathbf{X}\beta \rangle + \langle \mathbf{X}\beta, \mathbf{X}\beta \rangle. \end{aligned} \quad (5.25)$$

Since (5.25) holds for all vectors β , it certainly holds for $\beta = \hat{\beta}$. Thus

$$\begin{aligned} g(\hat{\beta}) &= \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta} \rangle + \langle \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta} \rangle \\ &= \langle \mathbf{y}, \mathbf{y} \rangle - \langle \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta} \rangle, \end{aligned} \quad (5.26)$$

and

$$\begin{aligned} g(\beta) - g(\hat{\beta}) &= \langle \mathbf{X}\beta, \mathbf{X}\beta \rangle - 2\langle \mathbf{X}\hat{\beta}, \mathbf{X}\beta \rangle + \langle \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta} \rangle \\ &= \langle \mathbf{X}\beta - \mathbf{X}\hat{\beta}, \mathbf{X}\beta - \mathbf{X}\hat{\beta} \rangle \\ &= \langle \mathbf{X}(\beta - \hat{\beta}), \mathbf{X}(\beta - \hat{\beta}) \rangle. \end{aligned} \quad (5.27)$$

Proof of Theorem 5.1. We first show that any solution to (5.21) minimizes $g(\beta)$. To see this, observe that

$$g(\beta) - g(\hat{\beta}) = \langle \mathbf{X}(\beta - \hat{\beta}), \mathbf{X}(\beta - \hat{\beta}) \rangle \geq 0, \quad (5.28)$$

so that $g(\beta) \geq g(\hat{\beta})$ for all vectors β . Thus $\hat{\beta}$ minimizes $g(\beta)$ and so is a MLE of β .

On the other hand, suppose that $\hat{\beta}_1$ minimizes $g(\beta)$. Then $g(\beta) \geq g(\hat{\beta}_1)$, so in particular $g(\hat{\beta}) \geq g(\hat{\beta}_1)$. Now using the identity (5.23) with $\beta = \hat{\beta}$ gives $g(\hat{\beta}_1) \geq g(\hat{\beta})$. Thus, $g(\hat{\beta}_1) = g(\hat{\beta})$, and using (5.23) again, we get

$$\langle \mathbf{X}(\hat{\beta}_1 - \hat{\beta}), \mathbf{X}(\hat{\beta}_1 - \hat{\beta}) \rangle = 0 \quad (5.29)$$

so that

$$\mathbf{X}(\hat{\beta}_1 - \hat{\beta}) = \mathbf{0}. \quad (5.30)$$

Multiplying both sides of (5.30) by \mathbf{X}^T on the left and using (5.21) again, we get

$$\mathbf{X}^T \mathbf{X} \hat{\beta}_1 - \mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{X} \hat{\beta}_1 - \mathbf{X}^T \mathbf{y} = \mathbf{0}. \quad (5.31)$$

Thus, $\mathbf{X}^T \mathbf{X} \hat{\beta}_1 = \mathbf{X}^T \mathbf{y}$ and $\hat{\beta}_1$ satisfies (5.21). Hence, every MLE of β is obtained by solving (5.21).

If $\mathbf{X}^T \mathbf{X}$ has an inverse, then the solution to (5.21) is unique, so that there is a unique MLE given by (5.22a). ■

Since Theorem 5.1 shows that the maximum likelihood estimator $\hat{\beta}$ of β can be obtained by minimizing the *residual sum of squares* $g(\beta) = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle$ with respect to β and $g(\beta)$ is differentiable, this minimum can be found by the standard calculus method of solving the simultaneous equations

$$\frac{\partial g(\beta)}{\partial \beta_j} = 0, \quad j = 0, 1, 2, \dots, m. \quad (5.32)$$

Carrying out these differentiations (see Exercise 5.5) we again arrive at the normal equations $\mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{y}$. When written out in full, these equations become

$$\begin{bmatrix} n & \sum_{k=1}^n x_{k1} & \cdots & \sum_{k=1}^n x_{km} \\ \cdot & \sum_{k=1}^n x_{k1}^2 & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \sum_{k=1}^n x_{km} & \cdot & \cdots & \sum_{k=1}^n x_{km}^2 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n y_k \\ \cdot \\ \cdot \\ \cdot \\ \sum_{k=1}^n x_{km} y_k \end{pmatrix}. \quad (5.33)$$

To see this, we write \mathbf{X} in partitioned form as

$$\mathbf{X} = [\mathbf{1} | \mathbf{x}_1 | \mathbf{x}_2 | \cdots | \mathbf{x}_m] \quad (5.34)$$

where $\mathbf{1} = (1, 1, \dots, 1)^T$, so that

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{1} \\ \mathbf{x}_1 \\ \cdot \\ \cdot \\ \mathbf{x}_m \end{bmatrix} \quad (5.35)$$

and using the rules for multiplying partitioned matrices

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} \mathbf{1} \\ \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{bmatrix} [\mathbf{1} | \mathbf{x}_1 | \mathbf{x}_2 | \cdots | \mathbf{x}_m] = \begin{bmatrix} n & \langle \mathbf{1}, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{1}, \mathbf{x}_m \rangle \\ \langle \mathbf{1}, \mathbf{x}_1 \rangle & \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{1}, \mathbf{x}_m \rangle & \langle \mathbf{x}_m, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{x}_m, \mathbf{x}_m \rangle \end{bmatrix}. \quad (5.36)$$

Similarly,

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} \mathbf{1} \\ \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \langle \mathbf{1}, \mathbf{y} \rangle \\ \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_m, \mathbf{y} \rangle \end{bmatrix} \quad (5.37)$$

and writing (5.33)-(5.37) in components gives (5.32).

Theorem 5.1 shows that a maximum likelihood estimator of $\boldsymbol{\beta}$ must be a solution to the *normal equations* (5.21) (also called the *least squares equations*). However, this does not guarantee that the MLE exists, unless we show one of two things:

- (i) Equation (5.21) always has at least one solution.
- (ii) There exists a vector $\hat{\boldsymbol{\beta}}$ minimizing $g(\boldsymbol{\beta})$.

Of course if $(\mathbf{X}^T \mathbf{X})^{-1}$ exists, we are done, and this is the case we shall consider in most of our work. However, for completeness, we will show that (i) is true even if this is not the case. For this we need the following fact from linear algebra, stated as Lemma 5.1 without proof.

Lemma 5.1 *The set of linear equations $\mathbf{Ax} = \mathbf{y}$ has a solution if and only if $\langle \mathbf{y}, \mathbf{z} \rangle = 0$ where $\mathbf{A}^T \mathbf{z} = 0$.*

Theorem 5.2 *The least-squares equations (5.21) always have at least one solution.*

Proof. From Lemma 5.1, we must show that $\langle \mathbf{X}^T \mathbf{y}, \mathbf{z} \rangle = 0$ for every \mathbf{z} such that $\mathbf{X}^T \mathbf{X} \mathbf{z} = 0$. Now if $\mathbf{X}^T \mathbf{X} \mathbf{z} = 0$, then $\langle \mathbf{X}^T \mathbf{X} \mathbf{z}, \mathbf{z} \rangle = \langle \mathbf{X} \mathbf{z}, \mathbf{X} \mathbf{z} \rangle = 0$, so that $\mathbf{X} \mathbf{z} = 0$. Thus,

$$\langle \mathbf{X}^T \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{y}, \mathbf{X} \mathbf{z} \rangle = 0 \quad (5.38)$$

and the theorem is proved. ■

It now follows from Theorems 5.1 and 5.2 that the MLE of $\boldsymbol{\beta}$ can be found by solving a set of $m+1$ linear equations in $m+1$ unknowns. Since there are relatively few situations where such solutions can be found analytically, in virtually all cases numerical techniques need to be used. This in itself is a problem of considerable complexity, since as pointed out in Chapter 4 well known techniques such as Gaussian elimination may run into difficulty [112].

As we shall see, experimental data frequently leads to $\mathbf{X}^T \mathbf{X}$ being highly ill-conditioned which can seriously affect the numerical accuracy of the computed value of $\hat{\beta}$. Hence, in recent years, regression calculations are often made using techniques such as Cholesky factorization the **QR** decomposition or the SVD. We shall return to this matter in Chapter 9.

Despite the fact that modern statistical packages, such as SAS, SPSS and MINITAB use these techniques and are generally believed to be reliable one should be cautious in uncritically using these “black boxes” without examining the output for possible computational glitches. In searching the internet recently, we came across numerous questions concerning numerical results which appeared to be in error. In general, it is wise to check for things such as the signs of $\hat{\beta}_i, 0 \leq i \leq m$, and their magnitude, since ill-conditioning can cause even sophisticated methods to fail. In this regard, one can perform a simple test of accuracy by adding small errors to the data and running the regression again and comparing it to your original output. Small changes in $\beta_i, 0 \leq i \leq m$, suggest reliability of the software.

As for simple linear regression, we can estimate β by minimizing $g(\beta)$ regardless of the nature of the errors. In this case $\hat{\beta}$ is called the *ordinary least squares estimate* (or just least squares estimate of β) and is often abbreviated to OLS. Currently, this is the most common method for estimating β in the GLM, although considerable research has taken place in the last two decades on alternative approaches [27, 87]. Some of these procedures are discussed in later chapters.

Since the normal equations always have at least one solution, it is important to be able to identify when this solution is unique, without necessarily having to form $\mathbf{X}^T \mathbf{X}$. It turns out that we need only examine the columns of \mathbf{X} .

Theorem 5.3 $\mathbf{X}^T \mathbf{X}$ is nonsingular if and only if the columns of \mathbf{X} are linearly independent. (In this case (5.16) is called a full rank model.)

Proof. We will first show that the linear independence of the columns of \mathbf{X} implies the nonsingularity of $\mathbf{X}^T \mathbf{X}$. Suppose this is not the case. Then by Theorem 4.2 there is a non-zero vector \mathbf{c} such that

$$\mathbf{X}^T \mathbf{X} \mathbf{c} = \mathbf{0}. \quad (5.39)$$

Thus, taking the dot product on both sides of (5.39) with \mathbf{c} we get

$$\langle \mathbf{c}, \mathbf{X}^T \mathbf{X} \mathbf{c} \rangle = \langle \mathbf{X} \mathbf{c}, \mathbf{X} \mathbf{c} \rangle = 0. \quad (5.40)$$

But (5.40) says that $\mathbf{X} \mathbf{c} = \mathbf{0}$ and writing this out in full shows that

$$\sum_{i=0}^m c_i \mathbf{x}_i = \mathbf{0} \quad (5.41)$$

where $\mathbf{c} = (c_1, c_2, \dots, c_m)^T$, and \mathbf{x}_i is now the i -th column of \mathbf{X} . This shows that the columns of \mathbf{X} are linearly dependent, which contradicts our assumption about \mathbf{X} . Thus $\mathbf{X}^T \mathbf{X}$ is nonsingular.

On the other hand, suppose that $\mathbf{X}^T \mathbf{X}$ is nonsingular but the columns of \mathbf{X} are linearly dependent. Thus there exists a non-zero vector \mathbf{c} such that

$$\mathbf{X} \mathbf{c} = \mathbf{0}. \quad (5.42)$$

Multiplying both sides of (5.42) by \mathbf{X}^T gives

$$\mathbf{X}^T \mathbf{X} \mathbf{c} = \mathbf{0} \quad (5.43)$$

and this shows that $\mathbf{X}^T \mathbf{X}$ is singular. Again we arrive at a contradiction and so the columns of \mathbf{X} must be linearly independent. ■

For some types of problems, linear independence of the columns is easily checked a priori, and ideally should be done before using a statistical package. However, most modern packages are reasonably “idiot-proof” and will inform the user if a singular matrix is encountered during the process of computation. In most cases the program will stop execution with the output of an appropriate error message.

Although it is certainly possible for $\mathbf{X}^T \mathbf{X}$ to be singular, what is a more frequent occurrence, and more difficult to remedy, is the possibility that the columns of \mathbf{X} may be “approximately” linearly independent. Roughly speaking this means that there is at least one column, say the j -th, which is an approximate linear combination of a subset of the remaining ones. That is

$$\mathbf{x}_j = \sum_{l=1}^k c_{il} \mathbf{x}_{i_l} + \boldsymbol{\delta} \quad (5.44)$$

where $\{i_l\}_{l=1}^k$ is a subset of $\{0, 1, 2, \dots, j-1, j+1, \dots, m\}$ and $\boldsymbol{\delta}$ is a “small” vector. (More precise definitions are given in Belsley, Kuh and Welsch (BKW) [8] and will be discussed in Chapter 9.) Relation (5.44) suggests that \mathbf{x}_j may be superfluous in explaining \mathbf{Y} once $\{\mathbf{x}_{i_l}\}_{l=1}^k$ have been included in the model. This problem of *multicollinearity*, as it is frequently called, can make it difficult to calculate $\hat{\boldsymbol{\beta}}$ numerically and to interpret the statistical significance of the regression coefficients. Numerical examples illustrating this problem will be given later in the Chapter.

5.3.2 Some Analytical and Numerical Solutions of the Normal Equations

As we have already seen, even for simple linear regression, practical applications of regression analysis require a computer in order to do the calculations. However, there are a number of cases where closed form solutions can be obtained to the normal equations. Although some of these solutions may be used infrequently as practical computing tools, they are often quite useful in a variety of theoretical analyses of the regression problem.

Analytical Solutions

As our first example of an analytical solution to the normal equations we will rederive the least squares estimators for the simple linear regression model.

Example 5.6 Since we have only one regressor in this case, we will let $x_1 = x$ to make our results consistent with the notation used in Chapter 3. Thus, (3.1) of Chapter

3 can be written in vector-matrix forms as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (5.45)$$

or equivalently as in (3.4)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (5.46)$$

The least squares equations then become

$$\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_m \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_m \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (5.47)$$

which after doing the indicated matrix multiplications become

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \quad (5.48)$$

or equivalently

$$\begin{cases} n\hat{\beta}_0 & + \hat{\beta}_1 \sum_{i=1}^n x_i & = \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i & + \hat{\beta}_1 \sum_{i=1}^n x_i^2 & = \sum_{i=1}^n x_i y_i \end{cases} \quad (5.49)$$

which are clearly the same as equations (3.11) and (3.14) in Chapter 3.

For these equations to have a unique solution the columns of \mathbf{X} must be linearly independent and for this to be true at least two elements of (x_1, x_2, \dots, x_n) must be different. This is equivalent to having $S_{xx} \neq 0$. In this case (5.49) can be solved as in Chapter 3 to yield (3.18)-(3.19).

Example 5.7 (Centered variables) Although the normal equations for the simple linear regression equation can be solved explicitly in a fairly simple form, analytical solutions obtained by standard algebraic methods such as Cramer's rule become increasingly complex as the number of variables increases. In this regard, we introduce a simple trick which reduces the dimensionality of the system of normal equations by one. In particular, a regression problem with three variables can be solved using only two, rather than three simultaneous equations. In that case, the analytical solution becomes as tractable as that for the simple regression model.

To proceed, let

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad 1 \leq j \leq m, \quad (5.50)$$

denote the mean of the j -th column of \mathbf{X} . Then by adding and subtracting \bar{x}_j and \bar{y} from both sides of (5.11), it becomes

$$Y_i - \bar{y} = \beta_0 - \bar{y} + \sum_{j=1}^m \beta_j \bar{x}_j + \sum_{j=1}^m (x_{ij} - \bar{x}_j) \beta_j + \varepsilon_i, \quad 1 \leq i \leq n. \quad (5.51)$$

In particular, if $m = 1$, then

$$\begin{aligned} Y_i - \bar{y} &= \beta_0 - \bar{y} + \beta_1 \bar{x} + (x_i - \bar{x}) \beta_1 + \varepsilon_i \\ &\equiv \beta_{0c} + (x_i - \bar{x}) \beta_1 + \varepsilon_i, \quad 1 \leq i \leq n \end{aligned} \quad (5.52)$$

where, as before, we denote x_{1i} by x_i .

In this case it is easy to find the least squares estimates of $(\beta_{0c}, \beta_{1c})^T$ and relate them to the least squares estimates of $(\beta_0, \beta_1)^T$ in the original model. If we let $y_{ic} = y_i - \bar{y}$, $x_{ic} = x_i - \bar{x}$, it follows from (3.19) in Chapter 3 that

$$\hat{\beta}_{0c} = \bar{y}_c - \bar{x}_c \hat{\beta}_{1c} = 0 \quad (5.53)$$

since $\bar{y}_c = \bar{x}_c = 0$. Also,

$$\hat{\beta}_{1c} = \frac{\sum_{i=1}^n y_{ic} (x_{ic} - \bar{x}_c)}{\sum_{i=1}^n (x_{ic} - \bar{x}_c)^2} = \frac{\sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\beta}_1. \quad (5.54)$$

Since $\beta_{0c} = \beta_0 - \bar{y} + \beta_1 \bar{x}$, we can obtain the least squares estimate of β_0 by setting

$$\hat{\beta}_0 = \hat{\beta}_{0c} + \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (5.55)$$

From these calculations we obtain the following procedure to obtain the least squares estimates of (β_0, β_1) in the simple linear regression model.

- (i) Form the centered variables: $y_{ic} = y_i - \bar{y}$ and $x_{ic} = x_i - \bar{x}$.
- (ii) Regress y_{ic} on x_{ic} to obtain the least squares estimate $\hat{\beta}_1$ of β_1 .
- (iii) The least squares estimate $\hat{\beta}_0$ of β_0 is then obtained from (5.55).

Using some tedious algebra resulting from the representation (5.49) we can obtain the least squares estimates of $\hat{\beta}$ in the GLM in a way similar to that for $m = 1$.

- (i') Find the least squares estimates of $(\beta_1, \beta_2, \dots, \beta_m)^T$ by solving the modified least squares equations

$$[\mathbf{X}_c^T \mathbf{X}_c] \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} = \mathbf{X}_c^T \mathbf{y}_c \quad (5.56)$$

where $[\mathbf{X}_c]_{ij} = [x_{ij} - \bar{x}_j]$, $1 \leq i \leq n$, $1 \leq j \leq m$, is the *centered design matrix* and

$$\mathbf{y}_c = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})^T \quad (5.57)$$

is the vector of *centered observations*.

(ii') $\hat{\beta}_0$ is then given by

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^m \hat{\beta}_j \bar{x}_j. \quad (5.58)$$

Note that this procedure requires solving m equations, one less than for the full least squares equations. In particular, the coefficient estimates for a two variable model may be obtained by solving only two equations in two unknowns, a task generally no more difficult than for the simple regression model. A numerical example illustrating these calculations follows.

Example 5.8 (Centered variables; $m = 2$) We begin by considering the general model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \quad (5.59)$$

By centering the variables we can rewrite (5.59) as

$$\begin{aligned} y - \bar{y} &= \beta_0 - \bar{y} + \beta_1 (x_1 - \bar{x}_1) + \beta_1 \bar{x}_1 + \beta_2 (x_2 - \bar{x}) + \beta_2 \bar{x}_2 + \varepsilon \\ &= \beta_0 - \bar{y} + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \beta_1 x_{1c} + \beta_2 x_{2c} + \varepsilon \end{aligned} \quad (5.60)$$

where $y_c = y - \bar{y}$, $x_{1c} = x_1 - \bar{x}_1$ and $x_{2c} = x_2 - \bar{x}_2$. In this form the model becomes

$$Y_c = \beta_0^* + \beta_1 x_{1c} + \beta_2 x_{2c} + \varepsilon \quad (5.61)$$

where

$$\beta_0^* = \beta_0 - \bar{y} + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2. \quad (5.62)$$

One can now estimate $(\beta_0^*, \beta_1, \beta_2)^T$ by least squares. As for the case of simple linear regression it can be shown that the least squares estimates of $(\beta_0^*, \beta_1, \beta_2)^T$ are

$$\hat{\beta}_0^* = 0, \quad \hat{\beta}_{1c} = \hat{\beta}_1, \quad \hat{\beta}_{2c} = \hat{\beta}_2 \quad (5.63)$$

where $\hat{\beta}_{ic}$ are the least squares estimates of β_{ic} , $i = 1, 2$ from (5.61) and $\hat{\beta}_i$, $i = 1, 2$, are the usual least squares estimates. Thus, it suffices to start with the model

$$Y_c = \beta_1 x_{1c} + \beta_2 x_{2c} + \varepsilon, \quad (5.64)$$

estimate $(\beta_1, \beta_2)^T$ by least squares and then estimate β_0 in (5.59) by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2. \quad (5.65)$$

This is a simplification because now only two simultaneous equations have to be solved to estimate $(\beta_0, \beta_1, \beta_2)^T$. To estimate $(\beta_1, \beta_2)^T$ we use (5.46) where

$$\mathbf{X}_c = \begin{bmatrix} x_{11c} & x_{12c} \\ x_{21c} & x_{22c} \\ \vdots & \vdots \\ x_{n1c} & x_{n2c} \end{bmatrix}, \quad (5.66)$$

so that (dropping the c 's for convenience)

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix}. \quad (5.67)$$

Introducing the shorthand notation $\sum_{i=1}^n x_{i1}^2 = \sum x_1^2$, $\sum_{i=1}^n x_{i1}x_{i2} = \sum x_1x_2$ etc., Eq. (5.46) can be solved using Cramer's rule to give

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{\begin{bmatrix} (\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y) \\ -(\sum x_1x_2)(\sum x_1y) + (\sum x_1^2)(\sum x_2y) \end{bmatrix}}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}. \quad (5.68)$$

Thus,

$$\hat{\beta}_1 = \frac{(\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}, \quad (5.69)$$

$$\hat{\beta}_2 = \frac{(\sum x_1^2)(\sum x_2y) - (\sum x_1x_2)(\sum x_1y)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}. \quad (5.70)$$

Using (5.69)-(5.70) we now consider fitting the model

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$$

to the following data:

x_1	x_2	y
-5	5	11
-4	4	11
-1	1	8
2	-3	2
2	-2	5
3	-2	5
3	-3	4
$\bar{x}_1 = 0$	$\bar{x}_2 = 0$	$\bar{y} = 46/7$

To do the fitting using (5.69)-(5.70) we first compute the centered data. This is given by

x_{1c}	x_{2c}	y_c
-5	5	31/7
-4	4	31/7
-1	1	10/7
2	-3	-32/7
2	-2	-11/7
3	-2	-11/7
3	-3	-18/7
$\sum x_{1c} = 0$	$\sum x_{2c} = 0$	$\sum y_{ic} = 0$

We now compute the various sums of squares and cross products required in Eqs. (5.68)-(5.70).

x_{1c}^2	x_{2c}^2	$x_{1c}x_{2c}$	$x_{1c}y$	$x_{2c}y$
25	25	-25	-155/7	155/7
16	16	-16	-124/7	124/7
1	1	-1	-10/7	10/7
4	9	-6	-64/7	96/7
4	4	-4	-22/7	22/7
9	4	-6	-33/7	22/7
9	9	-9	-54/7	54/7
<hr/>				
$\sum x_1^2 = 68$	$\sum x_2^2 = 68$	$\sum x_1x_2 = -67$	$\sum x_1y = -\frac{462}{7}$	$\sum x_2y = \frac{483}{7}$

From the data, (5.62) and (5.69)-(5.70) we get $\sum x_1^2 \sum x_2^2 - (\sum x_1x_2) = 68^2 - 67^2 = 135$, and

$$\begin{aligned}\hat{\beta}_1 &= \frac{-68(462/7) + 67(483/7)}{135} = 1, \\ \hat{\beta}_2 &= \frac{-67(462/7) + 68(483/7)}{135} = 2,\end{aligned}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 = 46/7.$$

Finally, we note that although using centered variables can simplify computations, its primary use these days is to improve the conditioning of the normal equations for better numerical accuracy and statistical interpretation.

Example 5.9 (Centered and scaled variables) Sometimes, in addition to centering the regressor variables, it is helpful to scale them as well. The most commonly used scaling is one in which the Euclidean length of each column of \mathbf{X}_c has length one. That is, if \mathbf{x}_j is the j -th column of \mathbf{X}_c , then each element of \mathbf{x}_j is divided by

$$s_j = \left[\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right]^{1/2}. \quad (5.71)$$

If \mathbf{X}_{sc} denotes the *centered and scaled design matrix*, then

$$\mathbf{X}_{sc} = \mathbf{X}_c \mathbf{S} \quad (5.72)$$

where \mathbf{S} is the diagonal matrix $\mathbf{S} = \text{diag}(1/s_1, 1/s_2, \dots, 1/s_m)$.

The centered and scaled version of the model is

$$\mathbf{Y}_c = \mathbf{X}_{sc}\boldsymbol{\beta}_s + \boldsymbol{\varepsilon}, \quad \mathbf{Y}_c = \mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1} \quad (5.73)$$

and the least squares estimate of $\boldsymbol{\beta}_s$ is given by

$$\hat{\boldsymbol{\beta}}_s = (\mathbf{X}_{sc}^T \mathbf{X}_{sc})^{-1} \mathbf{X}_{sc}^T \mathbf{y}_c. \quad (5.74)$$

Using (5.74) and the fact that the OLS estimates of $\hat{\beta}_i, 1 \leq i \leq m$, are obtained by solving the centered variable equations, it follows that

$$\hat{\beta}_c = \mathbf{S}\hat{\beta}_s. \quad (5.75)$$

To see (5.75) use (5.72) and (5.74) to give

$$\hat{\beta}_s = \left[(\mathbf{X}_c \mathbf{S})^T \mathbf{X}_c \mathbf{S} \right]^{-1} (\mathbf{X}_c \mathbf{S})^T \mathbf{y}_c \quad (5.76)$$

Since \mathbf{S} is diagonal, $\mathbf{S}^T = \mathbf{S}$, so that $(\mathbf{X}_c \mathbf{S})^T = \mathbf{S} \mathbf{X}_c^T$ and

$$\left[(\mathbf{X}_c \mathbf{S})^T \mathbf{X}_c \mathbf{S} \right]^{-1} = \left[\mathbf{S} \mathbf{X}_c^T \mathbf{X}_c \mathbf{S} \right]^{-1} = \mathbf{S}^{-1} (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{S}^{-1}. \quad (5.77)$$

Thus,

$$\hat{\beta}_s = \mathbf{S}^{-1} (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{S}^{-1} \mathbf{S} \mathbf{X}_c \mathbf{y}_c = \mathbf{S}^{-1} (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c \mathbf{y}_c = \mathbf{S}^{-1} \hat{\beta}_c. \quad (5.78)$$

Hence, $\hat{\beta}_c = \mathbf{S}\hat{\beta}_s$.

This leads to the following algorithm for computing $\hat{\beta}$ using centered and scaled variables:

- (i) Form \mathbf{X}_{sc} ;
- (ii) Estimate $\hat{\beta}_s$ from (5.74);
- (iii) Calculate $\hat{\beta}_i$ by multiplying the i -th component of $\hat{\beta}_s$ by $1/s_i, i = 1, 2, \dots, n$;
- (iv) Obtain $\hat{\beta}_0$ from (5.58).

Again, to reduce problems of ill-conditioning, many computer programs perform calculations using \mathbf{X}_{sc} . Because it follows from (5.71) and (5.72) that \mathbf{X}_{sc} is the matrix of *sample correlation coefficients* of the columns of \mathbf{X} one often says that the regression is done in *correlation form* [27, 87].

If one scales \mathbf{y}_c as well - denoted by \mathbf{y}_{sc} - the solution \mathbf{b} of

$$(\mathbf{X}_{sc}^T \mathbf{X}_{sc}) \mathbf{b} = \mathbf{X}_{sc} \mathbf{y}_{sc} \quad (5.79)$$

are sometimes referred to as *beta weights* [87].

Further discussion of using regression in correlation form will be given in Chapter 9.

Example 5.10 (Orthogonal variables) When \mathbf{X} has special structure it is often possible to obtain an analytic solution to the least squares equations. A particularly important example occurs when the columns $\mathbf{x}_j, 0 \leq j \leq m$ (for notational convenience, we denote the column of 1's in (5.15) by \mathbf{x}_0) are orthogonal; i.e., $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0, i \neq j$. Then using (5.36), $\mathbf{X}^T \mathbf{X}$ is the diagonal matrix $\text{diag}(\langle \mathbf{x}_0, \mathbf{x}_0 \rangle, \langle \mathbf{x}_1, \mathbf{x}_1 \rangle, \dots, \langle \mathbf{x}_m, \mathbf{x}_m \rangle)$. Then,

$$\begin{aligned} (\mathbf{X}^T \mathbf{X})^{-1} &= \text{diag} \left(\frac{1}{\langle \mathbf{x}_0, \mathbf{x}_0 \rangle}, \frac{1}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle}, \dots, \frac{1}{\langle \mathbf{x}_m, \mathbf{x}_m \rangle} \right) \\ &\equiv \text{diag}(\delta_0, \delta_1, \dots, \delta_m). \end{aligned} \quad (5.80)$$

Thus,

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \begin{bmatrix} \delta_0 & 0 & \cdots & 0 \\ 0 & \delta_1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \delta_m \end{bmatrix} \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \cdot \\ \cdot \\ \cdot \\ \langle \mathbf{x}_m, \mathbf{y} \rangle \end{pmatrix},\end{aligned}\quad (5.81)$$

so that

$$\hat{\beta}_i = \langle \mathbf{x}_i, \mathbf{y} \rangle \delta_i = \frac{\langle \mathbf{x}_i, \mathbf{y} \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle}, \quad 0 \leq i \leq m, \quad (5.82)$$

which requires only the formation of a number of inner products and no explicit solution of linear equations. For this and other reasons it is often desirable to obtain one's data so that the column's of \mathbf{X} are orthogonal (these are usually called *orthogonal designs*). This is frequently possible in the case of planned experiments, but is rarely the case for data that is collected outside of the laboratory. However, the possibility of "orthogonalizing" \mathbf{X} suggests a number of alternative approaches to solving the least square equations. When the orthogonalization is done using the Gram-Schmidt process this amounts to using the QR decomposition of the design matrix \mathbf{X} .

As a consequence, writing

$$\mathbf{X} = \mathbf{Q}\mathbf{R} \quad (5.83)$$

as in (4.217), the least squares estimate $\hat{\beta}$ of β in (5.16) is given by

$$\hat{\beta} = [(\mathbf{Q}\mathbf{R})^T \mathbf{Q}\mathbf{R}]^{-1} (\mathbf{Q}\mathbf{R})^T \mathbf{y} = (\mathbf{R}^T \mathbf{Q}^T \mathbf{Q}\mathbf{R})^{-1} \mathbf{Q}^T \mathbf{R}^T \mathbf{y} \quad (5.84)$$

Using the fact that \mathbf{Q} has orthogonal columns, $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n$ so that

$$\hat{\beta} = (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{Q}^T \mathbf{R}^T \mathbf{y} = (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{Q}^T \mathbf{z}, \quad \mathbf{z} = \mathbf{R}^T \mathbf{y} \quad (5.85)$$

Recalling that \mathbf{R} is the Cholesky factor of $\mathbf{X}^T \mathbf{X}$ we can obtain $\hat{\beta}$ by solving

$$(\mathbf{R}^T \mathbf{R}) \hat{\beta} = \mathbf{Q}^T \mathbf{z}. \quad (5.86)$$

Letting $\hat{\alpha} = \mathbf{R}\hat{\beta}$, $\hat{\beta}$ can be obtained by solving the two triangular systems

$$\mathbf{R}^T \hat{\alpha} = \mathbf{Q}^T \mathbf{z}, \quad \mathbf{R}\hat{\beta} = \hat{\alpha} \quad (5.87)$$

by forward and backward substitution. As noted in Chapter 4, since the QR decomposition can be obtained stably by repeatedly applying orthogonal matrices to \mathbf{X} , this procedure allows one to compute $\hat{\beta}$ without forming $\mathbf{X}^T \mathbf{X}$ or doing matrix inversions. As a consequence, this approach is one of the preferred computational methods for obtaining the least squares estimate $\hat{\beta}$.

Example 5.11 (The canonical form of the regression model) Another important form of the regression model using orthogonality is the *canonical form* (sometimes called the *principal components form*).

Since $\mathbf{X}^T \mathbf{X}$ is a symmetric matrix, by the spectral theorem there exists an orthogonal matrix \mathbf{Q} whose columns are eigenvectors of $\mathbf{X}^T \mathbf{X}$, i.e.,

$$\mathbf{Q}^T (\mathbf{X}^T \mathbf{X}) \mathbf{Q} = \mathbf{\Lambda} \quad (5.88)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_m)$ and $\lambda_i, 0 \leq i \leq m$, are the eigenvalues of $\mathbf{X}^T \mathbf{X}$. Letting $\mathbf{G} = \mathbf{XQ}$, then (5.88) becomes

$$\mathbf{G}^T \mathbf{G} = \mathbf{\Lambda} \quad (5.89)$$

where the columns of \mathbf{G} are given by

$$\boldsymbol{\gamma}_i = \mathbf{Xq}_i, \quad 0 \leq i \leq m \quad (5.90)$$

where $\mathbf{q}_i, 0 \leq i \leq m$, are the columns of \mathbf{Q} . From (5.89) $\mathbf{G}^T \mathbf{G}$ is the matrix $[\langle \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j \rangle], 0 \leq i, j \leq m$, which is diagonal because $\langle \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j \rangle = 0, i \neq j$ (i.e., the columns of \mathbf{G} are orthogonal). In fact,

$$\langle \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j \rangle = \langle \mathbf{Xq}_i, \mathbf{Xq}_j \rangle = \langle \mathbf{X}^T \mathbf{Xq}_i, \mathbf{q}_j \rangle = \lambda_i \langle \mathbf{q}_i, \mathbf{q}_j \rangle = 0, \quad i \neq j \quad (5.91)$$

because $\mathbf{q}_i, \mathbf{q}_j$ are eigenvectors of $\mathbf{X}^T \mathbf{X}$. Also,

$$\langle \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_i \rangle = \langle \mathbf{Xq}_i, \mathbf{Xq}_i \rangle = \langle \mathbf{X}^T \mathbf{Xq}_i, \mathbf{q}_i \rangle = \lambda_i \langle \mathbf{q}_i, \mathbf{q}_i \rangle = \lambda_i \quad (5.92)$$

since $\langle \mathbf{q}_i, \mathbf{q}_i \rangle = 1$.

The *canonical form* of the regression model is then given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{XQQ}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{G}\boldsymbol{\chi} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\chi} = \mathbf{Q}^T \boldsymbol{\beta} \quad (5.93)$$

and the least squares estimates of $\boldsymbol{\chi}_i$, are given by

$$\boldsymbol{\chi}_i = \langle \boldsymbol{\gamma}_i, \mathbf{y} \rangle / \lambda_i, \quad 0 \leq i \leq m, \quad (5.94)$$

as follows from (5.92). Hence,

$$\boldsymbol{\chi} = \mathbf{\Lambda}^{-1} \mathbf{G}^T \mathbf{y} \quad (5.95)$$

and it follows that

$$\hat{\boldsymbol{\beta}} = \mathbf{Q}\boldsymbol{\chi}. \quad (5.96)$$

To verify this, consider

$$\begin{aligned} \mathbf{X}^T \mathbf{XQ}\boldsymbol{\chi} &= \mathbf{X}^T \mathbf{XQ}\mathbf{\Lambda}^{-1} \mathbf{G}^T \mathbf{y} = \mathbf{X}^T \mathbf{XQ}\mathbf{\Lambda}^{-1} \mathbf{Q}^T \mathbf{X}^T \mathbf{y} \\ &= \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{y}, \end{aligned} \quad (5.97)$$

so $\hat{\boldsymbol{\beta}}$ is the unique solution to the normal equations. We will find this form of the regression particularly useful in Chapter 9.

5.3.3 Numerical Examples

We now turn our attention to giving a number of practical numerical examples of the use of multiple regression analysis. As for simple linear regression, the process of building a regression model is an iterative one. Usually one begins by using physical, economic or other knowledge to determine appropriate explanatory variables $\mathbf{x}_i, 1 \leq i \leq m$, for a

given measured response \mathbf{Y} . To determine if the response varies linearly with \mathbf{x}_i one can then make scatter plots of y against \mathbf{x}_i - generally these will be of similar appearance to those for simple linear regression. For those variables which appear to be useful predictors a linear model (5.11) is postulated and then fit (at least initially) by least squares. This is then followed by looking at goodness of fit, statistical significance, residual plots and other diagnostics to determine an adequate final model.

Example 5.12 (Housing data) To predict housing prices, the data in Table 5.1 were gathered. A scatter plot, Figure 5.2, of price y in dollars against square footage (x_1) shows a clear upward linear trend.

The plot of price against age (x_2) in years, Figure 5.3, while not as precise, shows a generally decreasing price as the age of a house increases. On the basis of these plots we consider a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

to estimate housing price as a function of square footage and age. The model was fit to the data in Table 5.1 and the coefficient estimates were

$$\hat{\beta}_0 = 13,239, \hat{\beta}_1 = 60.589, \hat{\beta}_2 = -1,726.8.$$

The fitted model is then

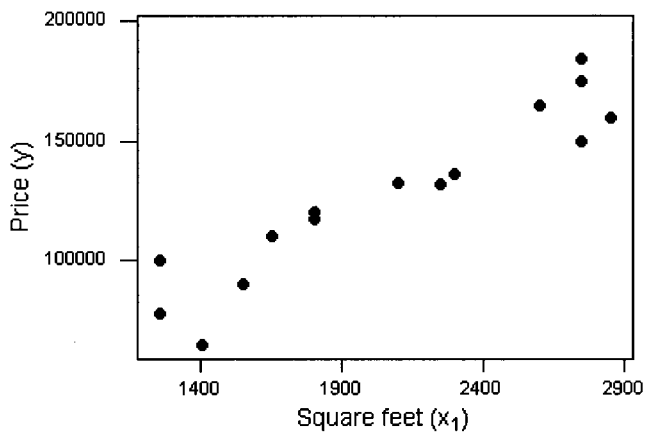
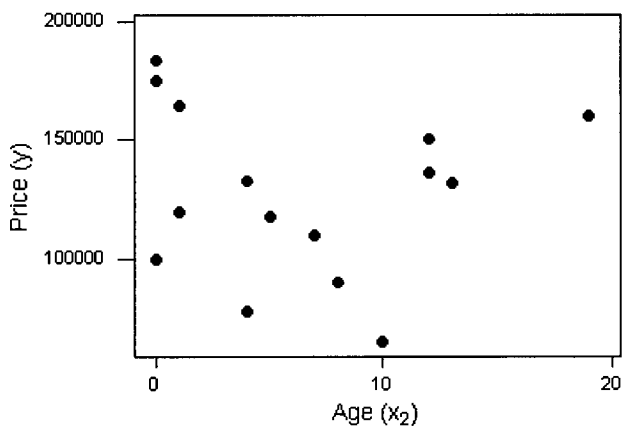
$$\hat{y} = 13,239 + 60.589x_1 - 1,726.8x_2. \quad (5.98)$$

From this we see that housing prices are approximately \$60.59 per square foot while the price of a house declines at about \$1,727 for each year of age. These figures are in accordance with the behavior of the scatter plots in Figures 5.2-5.3 and the three-dimensional plot in Figure 5.4.

Table 5.1 Housing Price Data

Obs. No.	Square ft (x_1)	Age (x_2)	Price (y)
1	1,800	1	120,000
2	1,650	7	110,000
3	2,750	12	150,000
4	1,550	8	90,000
5	2,750	0	175,000
6	1,400	10	65,000
7	1,250	4	78,000
8	1,250	0	100,000
9	2,250	13	131,500
10	2,300	12	136,100
11	2,750	0	184,000
12	2,600	1	164,500
13	2,850	19	160,000
14	2,100	4	132,500
15	1,800	5	117,500

Example 5.13 Although most authors in regression analysis generally recommend using scatter plots in the beginning stages of model building, they are not always reliable.

Figure 5.2: Scatter plot for price (y) versus square feet (x_1)Figure 5.3: Scatter plot of price (y) versus age (x_2)

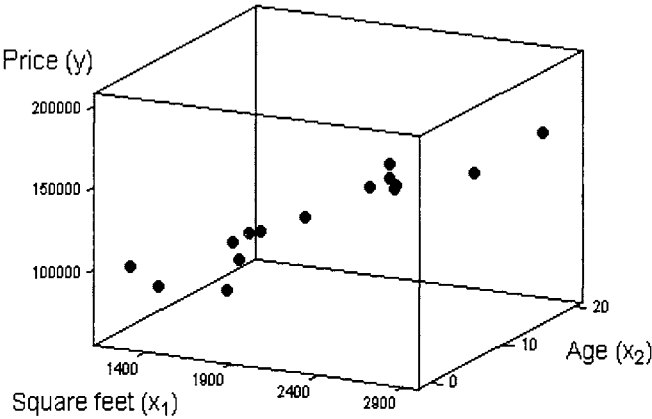


Figure 5.4: 3-D Scatter plot for Housing data - y versus x_1 and x_2

An interesting example is given by Montgomery, Peck and Vining [87] (See also Daniel and Wood [22]). They constructed scatter plots for the data given in Table 5.2.

Table 5.2 A Data Set		
x_1	x_2	y
2	1	10
3	2	17
4	5	48
1	2	27
5	6	55
6	4	26
7	3	9
8	4	16

The scatter plots of y against x_1 and x_2 are shown in Figures 5.5 and 5.6 respectively. Figure 5.5 shows a general random scatter while Figure 5.6 shows a general linear trend. To check these impressions we fitted separate least squares lines to each of these sets of data with the following results. For Figure 5.5

$$\begin{aligned}\hat{\beta}_0 &= 25.57, & t_0 &= 1.99 \\ \hat{\beta}_1 &= -0.571, & t_1 &= -0.20\end{aligned}$$

leading us to accept the hypothesis that $\beta_1 = 0$, while for Figure 5.6

$$\begin{aligned}\hat{\beta}_0 &= -1.34, & t_0 &= -0.14 \\ \hat{\beta}_1 &= 8.101, & t_1 &= 3.23\end{aligned}$$

leading us to accept the hypothesis that $\beta_1 \neq 0$.

Those results support our visual impression and might lead us to propose

$$Y = 25 + 8x_2 + \varepsilon$$

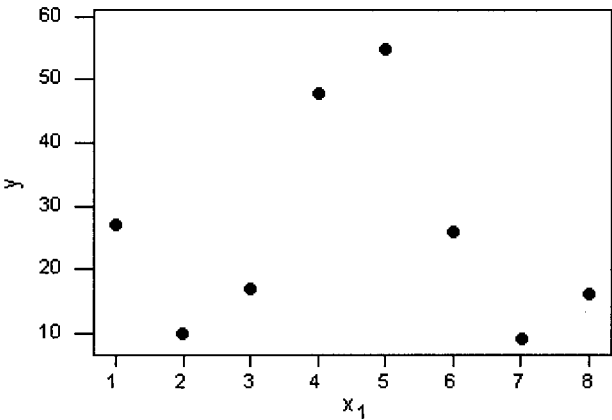


Figure 5.5: Scatter plot for y versus x_1 in Ex. 5.13.

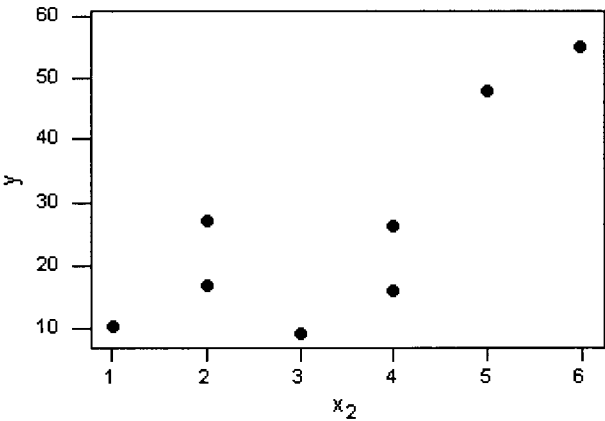


Figure 5.6: Scatter plot for y versus x_2 in Ex. 5.13.

as a reasonable model for this data. To check this we fitted the model

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$$

to the data. The results were

$$\hat{\beta}_0 = 8.00, \hat{\beta}_1 = -5.00, \hat{\beta}_2 = 12.00$$

and the fit was perfect since $R^2 = 1.00$ (see Section 5.6). This is not surprising in light of the fact that the data in Table 5.2 were obtained by choosing points on the plane

$$\hat{y} = 8 - 5x_1 + 12x_2$$

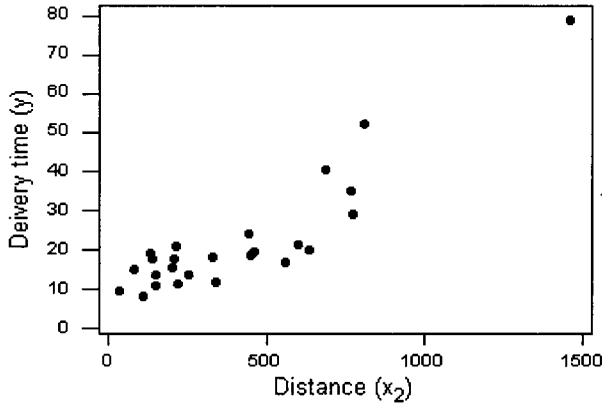
which is a totally different model than suggested by the one-dimensional scatter plots. Although these results indicate that scatter plots can be confusing for understanding multivariate data, we will continue to use them where appropriate. In Chapter 6 we will discuss other plots of multivariate data which have behavior similar to that of scatter plots of univariate data

Example 5.14 (Drink delivery data) In Examples 3.4 and 3.10 we examined the relation between the time to deliver drinks and the number of cases delivered. Although a linear model appeared reasonable. It was noted in Example 3.4 that observations 9 and 22 appeared to be outliers and there was some skew to the residuals. This suggests that perhaps other variables might be useful to improve the model. In this regard, further data was obtained on the distance (in feet) that the delivery person had to walk. That data are given in the “feet” column of Table 5.3 along with the data from Table 3.5.

Table 5.3 Drink Delivery data							
Obs. No.	Cases x_1	Feet x_2	Time y	Obs. No.	Cases x_1	Feet x_2	Time y
1	7	560	16.68	14	6	462	19.75
2	3	220	11.50	15	9	448	24.00
3	3	340	12.03	16	10	776	29.00
4	4	80	14.88	17	6	200	15.35
5	6	150	13.75	18	7	132	19.00
6	7	330	18.11	19	3	36	9.50
7	2	110	8.00	20	17	770	35.10
8	7	210	17.83	21	10	140	17.90
9	30	1460	79.24	22	26	810	52.32
10	5	605	21.50	23	9	450	18.75
11	16	688	40.33	24	8	635	19.83
12	10	215	21.00	25	4	150	10.75
13	4	255	13.50				

A scatter plot of time against distance shown in Figure 5.7 suggests a linear trend. This, in combination with our previous analysis, suggests a linear model

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon \tag{5.99}$$

Figure 5.7: Scatter plot of distance (y) versus time (x_2)

for this data. As a consequence the data in Table 5.3 were fit using least squares. The results were

$$\hat{\beta}_0 = 2.341, \hat{\beta}_1 = 1.6159, \hat{\beta}_2 = 0.014385.$$

One should note that the estimate of the slope for the effect of cases has changed substantially from its estimate 2.1762 when “distance” was excluded. This suggests that the estimate 2.1762 was biased by the omission of “distance.” We shall return to this in Chapters 6 and 8.

Example 5.15 (Birth weight data) To illustrate the use of a dummy variable in regression analysis we reconsider the birth weight data discussed in Examples 3.6 and 3.18. As noted in Example 3.6 Figure 3.9 shows that the residuals break up into two groups, the first underpredicting and the second overpredicting - a similar behavior is shown in Figure 3.9 where the residuals appear to fall along two parallel lines. Again, this suggests a missing variable. When the residuals are labeled by the sex of the child, it now appears that accounting for this might improve the model.

To do this we define a dummy variable

$$x_2 = \begin{cases} 1, & \text{if child is male,} \\ 0, & \text{if child is female.} \end{cases} \quad (5.100)$$

We then consider the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (5.101)$$

to account for the observations. From (5.101) we arrive at two models;

$$\begin{aligned} Y_M &= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon; & \text{for males,} \\ Y_F &= \beta_0 + \beta_1 x_1 + \varepsilon; & \text{for females,} \end{aligned} \quad (5.102)$$

so that β_2 is the difference between male and female weights for the same gestation age. Geometrically, (5.102) represents two parallel lines, with possibly different intercepts. To

account for possibly different slopes we include an *interaction term* $\beta_3 x_1 x_2$ as in Example 5.4 Then, our expanded model for the birth weight data is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underbrace{\beta_3 x_1 x_2}_{x_3} + \varepsilon. \quad (5.103)$$

This model was fit by least squares to the data in Table 3.7 where the first 12 children are males and the last 12 are females. The coefficients are given by

$$\hat{\beta}_0 = -2142, \hat{\beta}_1 = 130.4, \hat{\beta}_2 = 1042, \hat{\beta}_3 = -22.73.$$

Further analysis will be given later.

There are a number of data sets in the statistics literature which have become well known because of various difficulties associated with their analysis. One of these is the Hald data [49] given in Example 5.16. Another is the Longley data [76] discussed in Example 5.17.

Example 5.16 (Hald data) These data involve a study discussing the heat evolved in the hardening of cement. The variables are:

- y = heat evolved in calories/gram of cement,
- x_1 = percentage of tricalcium aluminate,
- x_2 = percentage of tricalcium silicate,
- x_3 = percentage of tetracalcium aluminoferrite,
- x_4 = percentage of dicalcium phosphate.

As one can easily verify, from Table 5.4 $x_1 + x_2 + x_3 + x_4 \simeq 100$ indicating that a possible collinearity exists among the predictor variables. The consequences of this will be investigated as we proceed.

Table 5.4 Hald's Cement data

Obs. No.	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

In [49] it was suggested that a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \quad (5.104)$$

would be appropriate to model the data. Hence (5.104) was fit using least squares giving

$$\begin{aligned}\hat{\beta}_0 &= 62.405, & \hat{\beta}_1 &= 1.5511, & \hat{\beta}_2 &= 0.5102, \\ \hat{\beta}_3 &= 0.019, & \text{and } \hat{\beta}_4 &= -0.1441.\end{aligned}$$

Example 5.17 (Longley data [76]) Suppose that we wish to determine the derived employment using regressor variables chosen from the Bureau of Labor Statistics. The variables are:

- y = Total derived employment (in thousands),
- x_1 = GNP implicit price deflator; 1954 = 100,
- x_2 = GNP, Gross National Product (in millions of dollars),
- x_3 = Unemployment (in thousands of persons),
- x_4 = Size of Armed Forces (in thousands),
- x_5 = Noninstitutional population 14 years of age and over (in thousands),
- x_6 = Time in years.

Table 5.5 Longley data

x_1	x_2	x_3	x_4	x_5	x_6	y
83.0	234,289	2,356	1,590	107,608	1947	60,323
88.5	259,426	2,325	1,456	108,632	1948	61,122
88.2	258,054	3,682	1,616	109,773	1949	60,171
89.5	284,599	3,351	1,650	110,929	1950	61,187
96.2	328,975	2,099	3,099	112,075	1951	63,221
98.1	346,999	1,932	3,594	113,270	1952	63,639
99.0	365,385	1,870	3,547	115,094	1953	64,989
100.0	363,112	3,578	3,350	116,219	1954	63,761
101.2	397,469	2,904	3,048	117,388	1955	66,019
104.6	419,180	2,822	2,857	118,734	1956	67,857
108.4	442,769	2,936	2,798	120,445	1957	68,169
110.8	444,546	4,681	2,637	121,950	1958	66,513
112.6	482,704	3,813	2,552	123,366	1959	68,655
114.2	502,601	3,931	2,514	125,368	1960	69,564
115.7	518,173	4,806	2,572	127,852	1961	69,331
116.9	554,894	4,007	2,827	130,081	1962	70,551

One considers that the following multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon \quad (5.105)$$

would be appropriate to fit the data. Hence (5.105) was used to find the least squares estimates, and the regression coefficients are

$$\begin{aligned}\hat{\beta}_0 &= -3482259, & \hat{\beta}_1 &= 15.1, & \hat{\beta}_2 &= -0.03582, & \hat{\beta}_3 &= -2.02, \\ \hat{\beta}_4 &= -1.03, & \hat{\beta}_5 &= -0.051, & \text{and } \hat{\beta}_6 &= 1829.2.\end{aligned}$$

As we see the nature of the problem is economic involving a price index, gross national product, unemployment and so on, so the explanatory variables seem to be naturally highly correlated.

5.3.4 Estimating σ^2

If the errors are $N(0, \sigma^2)$, then the MLE of σ^2 can be obtained by differentiating the likelihood function (5.17) with respect to σ^2 as for simple linear regression. Carrying this out (see Exercise 5.3) gives

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5.106)$$

where

$$\hat{y}_i = (\mathbf{X}\hat{\beta})_i, \quad 1 \leq i \leq n, \quad (5.107)$$

is the estimate of $E(Y_i)$, $1 \leq i \leq n$. Similarly, the MLE of σ is given by

$$\hat{\sigma}_{MLE} = \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]^{1/2}. \quad (5.108)$$

Since $y_i - \hat{y}_i = \hat{\varepsilon}_i$ is the i -th residual, (5.106) is just the average value of the sum of squares of the residuals. (As in Chapter 3, the sum of squares of the residuals $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is denoted by SSE .) In general, (5.106) is not used to estimate σ^2 since it is biased. Rather, the unbiased estimate

$$s^2 = \frac{SSE}{n - (m + 1)} \quad (5.109)$$

is the standard estimate of σ^2 and $s = \sqrt{s^2}$ is the customary estimate of σ (Note: s is generally a biased estimate of σ as was shown in Section 3.3.). In the more general case where the errors are uncorrelated and have constant variance σ^2 , but are not necessarily normal, (5.106) and (5.108) are also the standard estimates of σ^2 and σ respectively. Again the justification comes from the unbiasedness of s^2 as will be demonstrated in Section 5.4.

An alternative expression for SSE may be derived from Equation (5.20). Since $SSE = g(\hat{\beta})$ where $\hat{\beta}$ is the least squares estimate of β , we have

$$g(\hat{\beta}) = \langle \mathbf{y}, \mathbf{y} \rangle - \langle \mathbf{X}^T \hat{\beta}, \mathbf{X}^T \hat{\beta} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle - \langle \hat{\beta}, \mathbf{X}^T \mathbf{X} \hat{\beta} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle - \langle \hat{\beta}, \mathbf{X}^T \mathbf{y} \rangle. \quad (5.110)$$

Eq. (5.110) shows that SSE can be evaluated without forming the residuals. This may be computationally more efficient in some circumstances, since $\mathbf{X}^T \mathbf{y}$ will have already been obtained in order to solve the normal equations.

5.4 Properties of $(\hat{\beta}, s^2, \hat{\varepsilon})$

In this section we will establish a number of useful properties of the least squares estimator $\hat{\beta}$, the estimate of variance s^2 and the residual vector $\hat{\varepsilon}$. These generalize the results obtained in Chapter 3 for simple linear regression and are fundamental for the development of test procedures when the errors are normal. We first consider properties of $\hat{\beta}$ and s^2 .

Theorem 5.4 Let $\hat{\beta}$ be the least squares estimator of β in the full rank GLM. If the errors ε_i , $1 \leq i \leq n$, are independent with constant variance σ^2 , then

$$(i) E(\hat{\beta}) = \beta \text{ (}\hat{\beta} \text{ is unbiased);}$$

$$(ii) \Sigma(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1};$$

$$(iii) E(s^2) = \sigma^2 \text{ (} s^2 \text{ is unbiased);}$$

(iv) In addition, if ε_i , $1 \leq i \leq n$, are independent $N(0, \sigma^2)$ random variables, then $\hat{\beta}$ has a multivariate $N(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$ distribution. In particular, each $\hat{\beta}_i$ has a $N(\beta_i, \sigma^2 \delta_i)$ where δ_i is the i -th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$.

Proof. (i) From (5.22b) $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ so that

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{Y}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta. \end{aligned} \quad (5.111)$$

(ii) From (5.22b), $\hat{\beta} = \mathbf{A} \mathbf{Y}$ where $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, which shows that $\hat{\beta}$ is a linear combination of the Y_i 's. From Theorem 4.15 the variance-covariance matrix $\Sigma(\hat{\beta})$ of $\hat{\beta}$ is given by

$$\Sigma(\hat{\beta}) = \mathbf{A} \Sigma(\mathbf{Y}) \mathbf{A}^T. \quad (5.112)$$

But, $\Sigma(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$ and $\mathbf{A}^T = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$ so that

$$\Sigma(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}. \quad (5.113)$$

(iii) The proof that $E(s^2) = \sigma^2$ is a little more complicated than (i). We begin by obtaining an expression for the residual vector $\hat{\epsilon} = \mathbf{Y} - \mathbf{X} \hat{\beta}$.

Since $\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$

$$\hat{\epsilon} = \mathbf{Y} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = [\mathbf{I}_n - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \mathbf{Y} = (\mathbf{I}_n - \mathbf{H}) \mathbf{Y} \quad (5.114a)$$

where $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is called the *hat matrix*. Now note that $(\mathbf{I}_n - \mathbf{H}) \mathbf{X} = \mathbf{X} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$, so that

$$\begin{aligned} \hat{\epsilon} &= (\mathbf{I}_n - \mathbf{H}) \mathbf{Y} = (\mathbf{I}_n - \mathbf{H}) (\mathbf{X} \beta + \epsilon) \\ &= (\mathbf{I}_n - \mathbf{H}) \mathbf{X} \beta + (\mathbf{I}_n - \mathbf{H}) \epsilon = (\mathbf{I}_n - \mathbf{H}) \epsilon. \end{aligned} \quad (5.114b)$$

Now observe that $\mathbf{H}^T = \mathbf{H}$ and

$$\begin{aligned} \mathbf{H}^2 &= [\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] [\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \\ &= \mathbf{X} [(\mathbf{X}^T \mathbf{X}) \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \\ &= \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{H} \end{aligned} \quad (5.115)$$

and

$$(\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H}) \quad (5.116)$$

(so that \mathbf{H} is a symmetric idempotent matrix). Thus,

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \langle \hat{\varepsilon}, \hat{\varepsilon} \rangle = \langle (\mathbf{I}_n - \mathbf{H}) \varepsilon, (\mathbf{I}_n - \mathbf{H}) \varepsilon \rangle. \quad (5.117)$$

But,

$$\begin{aligned} \langle (\mathbf{I}_n - \mathbf{H}) \varepsilon, (\mathbf{I}_n - \mathbf{H}) \varepsilon \rangle &= \left\langle \varepsilon, (\mathbf{I}_n - \mathbf{H})^T (\mathbf{I}_n - \mathbf{H}) \varepsilon \right\rangle \\ &= \langle \varepsilon, (\mathbf{I}_n - \mathbf{H}) \varepsilon \rangle = \sum_{j=1}^n \sum_{i=1}^n g_{ij} \varepsilon_i \varepsilon_j \end{aligned} \quad (5.118)$$

where g_{ij} is the ij -th element of $\mathbf{I}_n - \mathbf{H}$.

Now taking the expectation of $\langle \varepsilon, (\mathbf{I}_n - \mathbf{H}) \varepsilon \rangle$ using the fact that $\varepsilon_i, 1 \leq i \leq n$, are uncorrelated gives

$$\begin{aligned} E[\langle \varepsilon, (\mathbf{I}_n - \mathbf{H}) \varepsilon \rangle] &= \sum_{j=1}^n \sum_{i=1}^n g_{ij} E(\varepsilon_i \varepsilon_j) \\ &= \sum_{j=1}^n \sum_{i=1}^n g_{ij} \text{Cov}(\varepsilon_i \varepsilon_j) \quad (\text{because } E(\varepsilon_i) = 0) \\ &= \sum_{i=1}^n g_{ii} \text{Var}(\varepsilon_i) = \sigma^2 \sum_{i=1}^n g_{ii}. \end{aligned} \quad (5.119)$$

We now need to evaluate $\sum_{i=1}^n g_{ii}$, which is the trace of $\mathbf{I}_n - \mathbf{H}$. Using the properties of the trace given in (4.31)-(4.32)

$$\begin{aligned} \text{tr}(\mathbf{I}_n - \mathbf{H}) &= n - \text{tr}[\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] = n - \text{tr}[(\mathbf{X}^T \mathbf{X})(\mathbf{X}^T \mathbf{X})^{-1}] \\ &= n - \text{tr}(\mathbf{I}_{m+1}) = n - (m + 1). \end{aligned} \quad (5.120)$$

Thus,

$$E(s^2) = E\left[\frac{SSE}{n - (m + 1)}\right] = \frac{[n - (m + 1)]\sigma^2}{n - (m + 1)} = \sigma^2. \quad (5.121)$$

(iv) Since each coefficient $\hat{\beta}_i, 1 \leq i \leq m + 1$ is a linear combination of the independent normal random variables $Y_i, 1 \leq i \leq n$, it then follows that $\hat{\beta}$ has a joint multivariate normal distribution. From (5.111) the mean vector is β and from (5.112) the variance-covariance matrix is $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$. Thus, $E(\hat{\beta}_i) = \beta_i$ and $\text{Var}(\hat{\beta}_i) = \sigma^2 \delta_i, 1 \leq i \leq m$. Hence $\hat{\beta}_i$ is a $N(\beta_i, \sigma^2 \delta_i)$ random variable. ■

We emphasize that properties (i)-(iii) of Theorem 5.4 are true even if $\varepsilon_i, 1 \leq i \leq n$, are only independent with common variance σ^2 . Normality is not needed. However, if the errors are normal, further important distributional properties of $\hat{\beta}$ and s^2 may be obtained. The proofs of some of these properties are somewhat technical and students

and instructors wishing to omit them may do so. However, these additional properties will be used throughout the text and so should be learned even if the proofs are not.

Theorem 5.5 *Let $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ be a full rank linear model where $\varepsilon_i, 1 \leq i \leq n$, are independent $N(0, \sigma^2)$. Then,*

- (i) $\hat{\beta}_i, 0 \leq i \leq m$, and s^2 are independent random variables;
- (ii) $(n - m - 1) s^2 / \sigma^2$ has a chi-square distribution with $n - m - 1$ degrees of freedom.
- (iii) If δ_i is the i -th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$, then

$$T_i = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{\delta_i}} \quad (5.122)$$

has a t distribution with $n - m - 1$ degrees of freedom.

(Note: Since $\text{Var}(\hat{\beta}_i) = \sigma^2 \delta_i$, $s\sqrt{\delta_i}$ is an estimate of $\sigma(\hat{\beta}_i)$. $s\sqrt{\delta_i}$ is usually called the standard error of $\hat{\beta}_i$. It will be denoted by $\hat{\sigma}(\hat{\beta}_i)$.)

- (iv) Let \mathbf{C} be an $r \times (m + 1)$ matrix of rank r . Then the quadratic form

$$\frac{q}{\sigma^2} = \frac{\left\langle \mathbf{C}(\hat{\beta} - \beta), [\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T]^{-1} \mathbf{C}(\hat{\beta} - \beta) \right\rangle}{\sigma^2} \quad (5.123)$$

has a chi-square distribution with r degrees of freedom.

Proof. (i) Consider

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \varepsilon) \\ &= \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \end{aligned} \quad (5.124)$$

so that

$$\hat{\beta} - \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon. \quad (5.125)$$

From (5.114b) $\hat{\varepsilon} = (\mathbf{I}_n - \mathbf{H})\varepsilon$ so the vector $(\hat{\beta} - \beta | \varepsilon)^T$ can be written as

$$\mathbf{Z} = \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\varepsilon} \end{pmatrix} = \begin{bmatrix} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ \mathbf{I}_n - \mathbf{H} \end{bmatrix} \varepsilon = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \varepsilon \quad (5.126)$$

where \mathbf{Z} is a function only of $\varepsilon_i, 1 \leq i \leq n$. Hence it follows that \mathbf{Z} has a degenerate multivariate normal distribution. The independence follows from Theorem 4.15 if we can show that $\text{Cov}(\hat{\beta}_i, \hat{\varepsilon}_j) = 0, 1 \leq i \leq m, 1 \leq j \leq n$. To do this we calculate $\Sigma(\mathbf{Z})$. Using (5.126) we find that

$$\begin{aligned} \Sigma(\mathbf{Z}) &= \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \Sigma(\varepsilon) \begin{bmatrix} \mathbf{A}^T & | & \mathbf{B}^T \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{A}^T & | & \mathbf{B}^T \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{A}^T & | & \mathbf{A}\mathbf{B}^T \\ \mathbf{B}\mathbf{A}^T & | & \mathbf{B}\mathbf{B}^T \end{bmatrix}. \end{aligned} \quad (5.127)$$

From the definition of \mathbf{Z} , the matrix $\mathbf{AB}^T = \left[\text{Cov}(\hat{\beta}_i, \hat{\varepsilon}_j) \right]$, $0 \leq i \leq m$, $1 \leq j \leq n$, so that

$$\begin{aligned} \mathbf{AB}^T &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{I}_n - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T - (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T - (\mathbf{X}^T \mathbf{X}) \mathbf{X}^T = \mathbf{0}. \end{aligned} \quad (5.128)$$

(ii) For this we observe that $(n - m - 1) s^2 / \sigma^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / \sigma^2 = \langle \boldsymbol{\varepsilon}, (\mathbf{I}_n - \mathbf{H}) \boldsymbol{\varepsilon} \rangle / \sigma^2$, which follows from (5.117). Since $\mathbf{I}_n - \mathbf{H}$ is symmetric and idempotent, it follows from the spectral theorem (Theorem 4.11) that

$$\mathbf{I}_n - \mathbf{H} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T \quad (5.129)$$

where \mathbf{Q} is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{I}_n - \mathbf{H}$ and $\boldsymbol{\Lambda}$ is an $n \times n$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{I}_n - \mathbf{H}$. Since the eigenvalues of $\mathbf{I}_n - \mathbf{H}$ are either 0 or 1, and $\text{tr}(\mathbf{I}_n - \mathbf{H}) = \text{rank}(\mathbf{I}_n - \mathbf{H}) = n - m - 1 = \text{tr}(\boldsymbol{\Lambda})$, $\boldsymbol{\Lambda}$ can be written in the form

$$\boldsymbol{\Lambda} = \left[\begin{array}{c|c} \mathbf{I}_{n-m-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right], \quad (5.130)$$

Hence,

$$\langle \boldsymbol{\varepsilon}, (\mathbf{I}_n - \mathbf{H}) \boldsymbol{\varepsilon} \rangle = \langle \boldsymbol{\varepsilon}, \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T \boldsymbol{\varepsilon} \rangle = \langle \mathbf{Q}^T \boldsymbol{\varepsilon}, \boldsymbol{\Lambda} \mathbf{Q}^T \boldsymbol{\varepsilon} \rangle. \quad (5.131)$$

Since \mathbf{Q}^T is an orthogonal matrix, letting $\mathbf{q} = \mathbf{Q}^T \boldsymbol{\varepsilon}$ it follows from Theorem 4.15 that \mathbf{q} is a vector of independent $N(0, \sigma^2)$ random variables. Thus,

$$\frac{\langle \mathbf{q}, \boldsymbol{\Lambda} \mathbf{q} \rangle}{\sigma^2} = \sum_{i=1}^{n-m-1} \frac{q_i^2}{\sigma^2} \quad (5.132)$$

is the sum of the squares of $n - m - 1$ independent $N(0, 1)$ random variables and so is χ^2 with $n - m - 1$ degrees of freedom.

(iii) From (iv) of Theorem 5.4 we know that $\hat{\beta}_i - \beta_i$ is $N(0, \sigma^2 \delta_i)$ so that $(\hat{\beta}_i - \beta_i) / \sigma \sqrt{\delta_i}$ is $N(0, 1)$. Now

$$T_i = \frac{(\hat{\beta}_i - \beta_i)}{s \sqrt{\delta_i}} = \frac{(\hat{\beta}_i - \beta_i) / \sigma \sqrt{\delta_i}}{s / \sigma} \quad (5.133)$$

and from (ii) s / σ is the square root of a $\chi^2(n - m - 1)$ random variable divided by its degrees of freedom. From (i) $(\hat{\beta}_i - \beta_i) / \sigma \sqrt{\delta_i}$ and s / σ are independent, so it follows from Section 2.8.3 that T_i has a t -distribution with $n - m - 1$ degrees of freedom.

(iv) To prove (5.123) let $\mathbf{W} = \mathbf{C} \hat{\boldsymbol{\beta}}$ and note that \mathbf{W} is $\mathbf{N}(\mathbf{C} \boldsymbol{\beta}, \sigma^2 \mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)$. Thus, $\mathbf{C} \hat{\boldsymbol{\beta}} - \mathbf{C} \boldsymbol{\beta}$ is $\mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)$, so we need only establish the following fact. Let \mathbf{z} be an r dimensional $\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ random vector, then

$$q = \langle \mathbf{z}, \boldsymbol{\Sigma}^{-1} \mathbf{z} \rangle / \sigma^2 \quad (5.134)$$

is $\chi^2(r)$.

To see this, we again use the spectral theorem to write

$$\Sigma = \mathbf{R}\mathbf{R}^T \quad (5.135)$$

where \mathbf{R} is nonsingular. Let $\mathbf{u} = \mathbf{R}^{-1}\mathbf{z}$, then $E(\mathbf{u}) = \mathbf{R}^{-1}E(\mathbf{z}) = \mathbf{0}$ and

$$\Sigma(\mathbf{u}) = \mathbf{R}^{-1}\Sigma(\mathbf{z})(\mathbf{R}^T)^{-1} = \mathbf{R}^{-1}\mathbf{R}\mathbf{R}^T(\mathbf{R}^T)^{-1} = \sigma^2\mathbf{I}_r. \quad (5.136)$$

Thus, the components of \mathbf{u} are independent $N(0, \sigma^2)$ random variables and

$$\begin{aligned} \frac{q}{\sigma^2} &= \frac{\langle \mathbf{z}, \Sigma^{-1}\mathbf{z} \rangle}{\sigma^2} = \frac{\langle \mathbf{R}\mathbf{u}, (\mathbf{R}^T)^{-1}\mathbf{R}^{-1}\mathbf{R}\mathbf{u} \rangle}{\sigma^2} \\ &= \frac{\langle \mathbf{R}^{-1}\mathbf{R}\mathbf{u}, \mathbf{R}^{-1}\mathbf{R}\mathbf{u} \rangle}{\sigma^2} = \frac{\langle \mathbf{u}, \mathbf{u} \rangle}{\sigma^2} = \sum_{i=1}^r \left(\frac{u_i}{\sigma}\right)^2. \end{aligned} \quad (5.137)$$

This shows that q/σ^2 is the sum of the squares of r independent $N(0, 1)$ random variables and so is $\chi^2(r)$. ■

5.4.1 Properties of $\hat{\varepsilon}$

Here we summarize a number of basic properties of the residual vector $\hat{\varepsilon}$. Some of these have already been discussed in Chapter 3 for the simple linear regression model and we will give suitable generalizations here.

Theorem 5.6 *Let $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ denote the residual vector from the least squares estimation of β in the GLM where the errors are uncorrelated with common variance σ^2 . Then,*

(i) $E(\hat{\varepsilon}) = \mathbf{0}$;

(ii) $\Sigma(\hat{\varepsilon}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$.

(iii) *If ε is $\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then $\hat{\varepsilon}$ has a $\mathbf{N}(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ distribution. If the model has an intercept, that is, $\beta_0 \neq 0$, then,*

(iv) $\sum_{i=1}^n \hat{\varepsilon}_i = 0$, and

(v) $\sum_{i=1}^n \hat{\varepsilon}_i \hat{y}_i = 0$.

Proof. (i) and (ii) and (iii) follow easily from the representation $\hat{\varepsilon} = (\mathbf{I}_n - \mathbf{H})\varepsilon$ given by (5.114b) and are left to the reader.

(iv) From Exercise 5.7 or by examining the first of the normal equations

$$\sum_{i=1}^n \left[y_i - \left(\hat{\beta}_0 + \sum_{j=1}^m x_{ij}\hat{\beta}_j \right) \right] = 0, \quad (5.138)$$

where the term in the parentheses is \hat{y}_i . Thus, $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$ as required.

(v) For this we observe that the normal equations can be written as

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}, \quad (5.139)$$

and taking the dot product of (5.139) with $\hat{\beta}$ gives

$$\langle \hat{\beta}, \mathbf{X}^T (\mathbf{y} - \mathbf{X}\hat{\beta}) \rangle = \langle \mathbf{X}\hat{\beta}, \mathbf{y} - \mathbf{X}\hat{\beta} \rangle = \langle \hat{\mathbf{y}}, \hat{\boldsymbol{\varepsilon}} \rangle = 0. \quad (5.140)$$

But, $\langle \hat{\mathbf{y}}, \hat{\boldsymbol{\varepsilon}} \rangle = \sum_{i=1}^n \hat{\varepsilon}_i \hat{y}_i$ and (v) follows. (Note that (v) does not require the model to have an intercept.) ■

Using (iv) and (v) of Theorem 5.6 we can obtain the following decomposition of the adjusted sum of squares for a model where $\beta_0 \neq 0$.

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SSE} \quad (5.141)$$

where, as in Chapter 3, *SSR* is the *regression sum of squares*. The proof of (5.141), as in Eq. (3.141) requires only that $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ and $\sum_{i=1}^n \hat{\varepsilon}_i \hat{y}_i = 0$ and so needs no further justification. As for simple linear regression, we will use this identity to introduce a generalization of R^2 in Section 5.6.

5.4.2 Further properties of $\Sigma(\hat{\beta})$

Because the variance-covariance matrix $\Sigma(\hat{\beta})$ of $\hat{\beta}$ plays such an important role in regression analysis we will now present several examples of its computation. For most numerical examples, this needs to be done with a computer. However, our emphasis here will be on the development of a number of alternative expressions for $\Sigma(\hat{\beta})$ arising primarily from various special forms of the regression model.

Example 5.18 ($\Sigma(\hat{\beta})$ for the simple linear regression model) To illustrate some of the power of matrix methods in doing regression calculations we will rederive the expressions for the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ in the simple linear regression model.

We note from Example 5.8 that in this case

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad (5.142)$$

and using Cramer's rule to find the inverse of a 2×2 matrix we find that

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{\det(\mathbf{X}^T \mathbf{X})} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \quad (5.143)$$

where

$$\begin{aligned} \det(\mathbf{X}^T \mathbf{X}) &= n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = n \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \\ &= n \sum_{i=1}^n (x_i - \bar{x})^2 = nS_{xx}. \end{aligned} \quad (5.144)$$

Thus,

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}. \quad (5.145)$$

From Theorem 5.4 we find that $\sigma^2(\hat{\beta}_i)$, $i = 0, 1$, are given by $\sigma^2\delta_i$, $i = 0, 1$ where δ_i , $i = 0, 1$ are the i -th diagonal elements of $(\mathbf{X}^T \mathbf{X})^{-1}$. Thus,

$$\sigma^2(\hat{\beta}_0) = \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n x_i^2 \quad (5.146)$$

and

$$\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \quad (5.147)$$

From (5.147) we see that the expression for $\sigma^2(\hat{\beta}_1)$ agrees with that given in Theorem 3.1 while that for $\sigma^2(\hat{\beta}_0)$ appears somewhat different. A little algebra, however shows that they agree. For this we observe that in Theorem 3.1 we found that

$$\sigma^2(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]. \quad (5.148)$$

But,

$$\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} = \frac{S_{xx} + n\bar{x}^2}{nS_{xx}} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}} = \sigma^2(\hat{\beta}_0) \quad (5.149)$$

as follows from the previous paragraph, so the two expressions for $\sigma^2(\hat{\beta}_0)$ agree.

In addition, we note from (5.143) that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \sum_{i=1}^n x_i / nS_{xx} = -\sigma^2 \bar{x} / S_{xx}$ which agrees with the expression found in Chapter 3 (and by a somewhat less tedious calculation).

Example 5.19 (Spectral analysis of $\sigma^2(\hat{\beta}_i)$) For some purposes, particularly for the analysis of multicollinearity, it is useful to see how the variances of $\hat{\beta}_i$ depend on the eigenvalues of $(\mathbf{X}^T \mathbf{X})^{-1}$.

For this we use the spectral theorem, so that $\mathbf{X}^T \mathbf{X} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$ and $(\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^T$. Now letting $\mathbf{q}_j = [q_{ij}]$ denote the j -th column of \mathbf{Q} , then

$$\begin{aligned} \mathbf{Q} \mathbf{\Lambda}^{-1} &= \begin{bmatrix} \mathbf{q}_0 & \mathbf{q}_1 & \dots & \mathbf{q}_m \\ \lambda_0 & \lambda_1 & & \lambda_m \end{bmatrix} \\ &= \left[\frac{q_{ij}}{\lambda_i} \right], \quad 0 \leq i, j \leq m, \end{aligned} \quad (5.150)$$

and $(\mathbf{Q}^T)_{ij} = q_{ji}$, $0 \leq i, j \leq m$. Thus the i -th diagonal element of $\mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^T$ is given by

$$\sum_{k=0}^m \left(\frac{q_{ik}}{\lambda_k} \right) (\mathbf{Q}^T)_{ki} = \sum_{k=0}^m \left(\frac{q_{ik}}{\lambda_k} \right) q_{ik} \quad (5.151)$$

so that

$$(\mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}^T)_{ii} = \sum_{k=0}^m \frac{q_{ik}^2}{\lambda_k}. \quad (5.152)$$

so that

$$[(\mathbf{X}^T\mathbf{X})^{-1}]_{ii} = \sum_{k=0}^m \frac{q_{ik}^2}{\lambda_i} \quad (5.153)$$

and

$$\sigma^2(\hat{\beta}_i) = \sigma^2 \sum_{k=0}^m \frac{q_{ik}^2}{\lambda_i}. \quad (5.154)$$

From (5.153) we see that if any of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ is very small, then the variances of all the $\hat{\beta}_i$'s may be large and so are estimated with poor precision. Since, as we will show in Chapter 9, it is the existence of small eigenvalues of $\mathbf{X}^T\mathbf{X}$ that characterizes multicollinearity, then large values of δ_i are suggestive of multicollinearity.

In Chapter 9 we will argue that when the regression is done in *correlation form* that values of $\delta_i \geq 10$ (in this case δ_i is called the *variance inflation factor* (VIF)) are suggestive of possible problems due to multicollinearity. Since most known computer packages give the VIFs as standard output, for now, we will monitor collinearity by examining these quantities. In Table 5.6 we list the VIFs for the data discussed in Examples 5.12-5.17.

Table 5.6 Variance Inflation Factors (VIFs) for Examples 5.12-5.17

Data in Examples	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
Housing prices data (Ex. 5.12)	1.0	1.0	-	-	-	-
Drink delivery data (Ex. 5.14)	3.1	3.1	-	-	-	-
Birth weight data (Ex. 5.15) x_3 in	2.3	449.7	449.7	-	-	-
Birth weight data (Ex. 5.15) x_3 out	1.0	1.0	-	-	-	-
Hald data (Ex. 5.16)	38.5	254.4	46.9	282.5	-	-
Longley data (Ex. 5.17)	135.5	1788.5	33.6	3.6	399.2	759.0

From Table 5.6 we see that both the Longley and Hald data have a number of large VIFs so we anticipate difficulty in interpreting the models proposed in Examples 5.12 and 5.14. The housing and drink delivery data appear to be free of multicollinearity problems while the birth weight data appears contradictory. (Can you suggest what the problem is?) Further discussion of this matter will be given in Chapter 7.

As we shall see, the presence of this phenomenon can confound the statistical interpretation of the estimated coefficients $\hat{\beta}_i$. As we proceed with our discussion of estimation and testing the reader should keep this relation in mind when confronting apparently contradictory results of a particular analysis.

Another useful result, derivable from (5.113) is an expression for the *total variance* of $\hat{\beta}$, $\sum_{i=0}^m \sigma^2(\hat{\beta}_i) \equiv \text{Var}(\hat{\beta})$. Since $\sigma^2(\hat{\beta}_i)$ is $\sigma^2\delta_i$, then

$$\text{Var}(\hat{\beta}) = \sigma^2 \sum_{i=0}^m \delta_i. \quad (5.155)$$

But $\sum_{i=0}^m \delta_i = \text{tr}[(\mathbf{X}^T \mathbf{X})^{-1}]$. From (4.141) $\text{tr}[(\mathbf{X}^T \mathbf{X})^{-1}] = \text{tr}(\mathbf{\Lambda}^{-1}) = \sum_{i=0}^m 1/\lambda_i$ so that

$$\text{Var}(\hat{\beta}) = \sum_{i=0}^m \sigma^2 / \lambda_i. \quad (5.156)$$

Example 5.20 ($\Sigma(\hat{\beta})$ for orthogonal models) When the model is orthogonal $\Sigma(\hat{\beta})$ is particularly easy to obtain. In this case, since the columns of \mathbf{X} are orthogonal

$$\mathbf{X}^T \mathbf{X} = \text{diag}[\langle \mathbf{x}_i, \mathbf{x}_i \rangle] \quad (5.157)$$

where \mathbf{x}_i is the i -th column of \mathbf{X} . Thus,

$$\Sigma(\hat{\beta}) = \sigma^2 \text{diag} \left[\frac{1}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle} \right] \quad (5.158)$$

and from this we find that $\hat{\beta}_i$ are uncorrelated and so in the normal model are independent and

$$\sigma^2(\hat{\beta}_i) = \frac{\sigma^2}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle}, \quad 0 \leq i \leq m. \quad (5.159)$$

which can be obtained without any matrix inversion.

Example 5.21 (Computing $\Sigma(\hat{\beta})$ from the centered and scaled models) Because one often uses the centered and/or the centered and scaled forms of \mathbf{X} to do calculations, it is important to see how one can obtain $\Sigma(\hat{\beta})$ in terms of the quantities \mathbf{X}_c and \mathbf{X}_{sc} . We begin with the centered model first by noting that

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} \gamma & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix} \quad (5.160)$$

where

$$\begin{aligned} \gamma &= (-\mathbf{B}\boldsymbol{\alpha}^T + 1)/n, & \mathbf{B} &= -\boldsymbol{\alpha}(\mathbf{X}_c^T \mathbf{X}_c)^{-1}/n, \\ \mathbf{D} &= (\mathbf{X}_c^T \mathbf{X}_c)^{-1}, & \boldsymbol{\alpha} &= \mathbf{1}\mathbf{X}_v, \end{aligned}$$

where $\mathbf{1} = (1, 1, \dots, 1)$ is an n -vector, and \mathbf{X}_v is the matrix of columns 1 to m of \mathbf{X} .

Thus, $\sigma^2(\hat{\beta}_i)$, $1 \leq i \leq m$, is given by

$$\sigma^2(\hat{\beta}_i) = \sigma^2 \left[(\mathbf{X}_c^T \mathbf{X}_c)^{-1} \right]_{ii}, \quad 1 \leq i \leq m, \quad (5.161)$$

and

$$\sigma^2(\beta_0) = \gamma. \quad (5.162)$$

We note that these can be computed in terms of quantities determined in the course of solving the centered normal equations.

In the case of centered and scaled variables we have from Example 5.9 that $\mathbf{X}_c = \mathbf{X}_{sc}\mathbf{S}^{-1}$ so that $\mathbf{X}_c^T \mathbf{X}_c = \mathbf{S}^{-1} \mathbf{X}_{sc}^T \mathbf{X}_{sc} \mathbf{S}^{-1}$ so that $(\mathbf{X}_c^T \mathbf{X}_c)^{-1} = \mathbf{S}(\mathbf{X}_{sc} \mathbf{X}_{sc})^{-1} \mathbf{S}$. Then $\sigma^2(\beta_i)$, $0 \leq i \leq m$, can be obtained from (5.161)-(5.162).

5.4.3 A Summary of OLS Estimators

Because the properties of $\hat{\beta}$, $\hat{\varepsilon}$ and s^2 that we have proved in the previous subsection are so important for the further development of regression analysis we feel it is worthwhile to gather them here in Table 5.7 for easy future reference.

In particular, those readers who have skipped the proof of Theorem 5.4 may simply refer to this section for the needed results as they arise in further work. We assume that we are considering the full rank GLM in the form (5.16) where the errors are uncorrelated with common variance σ^2 . For distributional properties we make the additional assumption that $\varepsilon_i \sim N(0, \sigma^2)$, $1 \leq i \leq n$.

Table 5.7 Summary of Properties of Ordinary Least Squares Estimators

Quantity	Properties
$\hat{\beta}$ (least squares estimator of β)	$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ $E(\hat{\beta}) = \beta$ $\Sigma(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, $Var(\hat{\beta}_i) = \sigma^2 \delta_i$, where δ_i is the i -th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$ $\hat{\beta}_i \sim N(\beta_i, \sigma^2 \delta_i)$
$\hat{\mathbf{Y}}$ (point estimator of $E(\mathbf{Y})$)	$\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta}$ $E(\hat{\mathbf{Y}}) = E(\mathbf{Y}) = \mathbf{X} \beta$ $\Sigma(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{H}$, $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, $Var(\hat{Y}_i) = \sigma^2 h_{ii}$, where h_{ii} is the i -th diagonal element of \mathbf{H} $\hat{Y}_i \sim N(\mu_i, \sigma^2 h_{ii})$
$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ (residual vector)	$\hat{\varepsilon} = (\mathbf{I}_n - \mathbf{H}) \varepsilon$ $E(\hat{\varepsilon}) = 0$ $\Sigma(\hat{\varepsilon}) = \sigma^2 (\mathbf{I}_n - \mathbf{H})$ $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ ($\beta_0 \neq 0$) $\sum_{i=1}^n \hat{\varepsilon}_i \hat{y}_i = 0$ $\hat{\varepsilon}_i \sim N(0, \sigma^2 (1 - h_{ii}))$
s^2 (estimator of σ^2)	$s^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 / (n - m - 1) = \langle \hat{\varepsilon}, \hat{\varepsilon} \rangle / (n - m - 1)$ $E(s^2) = \sigma^2$ $(n - m - 1) s^2 / \sigma^2$ is $\chi^2(n - m - 1)$ s^2 is independent of $\hat{\beta}_i, i = 0, 1, \dots, m$ $\hat{\sigma}(\hat{\beta}_i) = s \sqrt{\delta_i}$ is the <i>standard error</i> of $\hat{\beta}_i$ $T_i = \frac{\hat{\beta}_i - \beta_i}{s \sqrt{\delta_i}}$ has a t -distribution with $df = n - m - 1$
\mathbf{H} (the hat matrix)	$\hat{\mathbf{Y}} = \mathbf{H} \mathbf{Y}$ $\mathbf{H}^T = \mathbf{H}$, $(\mathbf{I}_n - \mathbf{H})^T = \mathbf{I}_n - \mathbf{H}$ (symmetry) $\mathbf{H}^2 = \mathbf{H}$, $(\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n - \mathbf{H}$ (idempotency) $\mathbf{H} \mathbf{X} = \mathbf{X}$ $\text{tr}(\mathbf{H}) = \sum_{i=1}^n h_{ii} = m + 1$, $1/n \leq h_{ii} \leq 1$ $\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = (\mathbf{I}_n - \mathbf{H})\mathbf{H} = \mathbf{0}$

5.5 The Gauss-Markov Theorem

As we have pointed out on several occasions, the least squares estimator of β in the GLM coincides with the MLE when the errors are independent and $N(0, \sigma^2)$. Since the least squares estimator is known to be efficient in this case, that is, it is the *minimum variance unbiased estimator* (MVUE) of β , it is interesting to determine what optimality properties are present if the errors are not normal.

As for simple linear regression the least squares estimator of β is the minimum variance unbiased linear estimator if only the errors are uncorrelated and have constant variance. This is the celebrated *Gauss-Markov theorem* and it is frequently invoked to justify the use of the least squares estimator even when the errors are not normal. Even though the least squares estimator is the “best linear unbiased estimator” (BLUE) there may be better linear biased estimators or even better nonlinear estimators. Some of these are discussed in [66, 87]. Notwithstanding, the use of these other estimators is not as widespread as the least squares estimator and their properties are not as well developed.

Definition 5.1 Let β be the vector of regression coefficients in the GLM. We say that $\hat{\beta}$ is a *linear estimator* of β if each β_i is a linear combination of the observations $Y_i, 1 \leq i \leq n$, so that

$$\hat{\beta}_i = \sum_{j=1}^n a_{ij} Y_j, \quad 0 \leq i \leq m. \quad (5.163)$$

If $\mathbf{A} = [a_{ij}]$ then (5.163) can be written as

$$\hat{\beta} = \mathbf{A}\mathbf{Y}. \quad (5.164)$$

Observe that if the model (5.16) has full rank, then the least squares estimator is linear, since

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (5.165)$$

and $\hat{\beta} = \mathbf{A}\mathbf{Y}$, where $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

Theorem 5.7 (Gauss-Markov theorem) *Consider the full rank GLM $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon_i, 1 \leq i \leq n$, are uncorrelated and have common variance σ^2 . Then the least squares estimator $\hat{\beta}$ of β is the minimum variance unbiased linear estimator of β . ($\hat{\beta}$ is often referred to as the BLUE estimator.)*

Proof. Let $\hat{\beta} = \mathbf{A}\mathbf{Y}$ be a linear estimator of β , then in order that $\hat{\beta}$ be unbiased we must have $E(\hat{\beta}) = \beta$ and this gives

$$E(\hat{\beta}) = \mathbf{A}E(\mathbf{Y}) = \mathbf{A}\mathbf{X}\beta = \beta. \quad (5.166)$$

Thus, $(\mathbf{A}\mathbf{X} - \mathbf{I}_m)\beta = \mathbf{0}$. Since this must hold for all vectors $\beta \in \mathbb{R}^{m+1}$, $\mathbf{A}\mathbf{X} - \mathbf{I}_{m+1} = \mathbf{0}$, or equivalently,

$$\mathbf{A}\mathbf{X} = \mathbf{I}_{m+1}. \quad (5.167)$$

We now compute the variance-covariance matrix of $\hat{\beta}$. From (5.112)

$$\Sigma(\hat{\beta}) = \mathbf{A}\Sigma(\mathbf{Y})\mathbf{A}^T. \quad (5.168)$$

and by assumption $\Sigma(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, so that

$$\Sigma(\hat{\beta}) = \sigma^2 \mathbf{A} \mathbf{A}^T. \quad (5.169)$$

Now write

$$\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D} \quad (5.170)$$

so that \mathbf{D} represents the difference between \mathbf{A} and the matrix for the least squares estimator. We will have proved the theorem if we can show $\mathbf{D} = \mathbf{0}$.

Substituting this expression for \mathbf{A} into (5.169) we get

$$\begin{aligned} \Sigma(\hat{\beta}) &= \sigma^2 \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D} \right] \left[\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{D}^T \right] \\ &= \sigma^2 \left[(\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{D} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}^T + \mathbf{D} \mathbf{D}^T \right]. \end{aligned} \quad (5.171)$$

Similarly, using the unbiasedness condition,

$$\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D} \right] \mathbf{X} = \mathbf{I}_{m+1} + \mathbf{D} \mathbf{X} = \mathbf{I}_{m+1} \quad (5.172)$$

so that $\mathbf{D} \mathbf{X} = \mathbf{0}$, and taking the transpose of this, $\mathbf{X}^T \mathbf{D}^T = \mathbf{0}$. Using these two conditions in (5.171) shows that

$$\Sigma(\hat{\beta}) = \sigma^2 \left[(\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{D} \mathbf{D}^T \right]. \quad (5.173)$$

Now examining the diagonal elements on the both sides of (5.173) gives

$$\text{Var}(\hat{\beta}_i) = \sigma^2 \left[\delta_i + \sum_{j=0}^n d_{ij}^2 \right]. \quad (5.174)$$

Since $\delta_i > 0$ and $\sum_{j=0}^n d_{ij}^2 \geq 0$, the variance of $\hat{\beta}_i$ is minimized by setting $\sum_{j=0}^n d_{ij}^2 = 0$. Thus $d_{ij} = 0, 1 \leq i \leq n$, and since this holds for all $j = 0, 1, \dots, m$, we find that $\mathbf{D} = \mathbf{0}$. Thus the linear unbiased estimator $\hat{\beta}$ which minimizes $\text{Var}(\hat{\beta}_i), 0 \leq i \leq m$, is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ which is the least squares estimator. ■

5.6 Testing the Fit - the Basic ANOVA Table

5.6.1 The Overall F -Test

As in the case of simple linear regression, we must regard any model as tentative, and having proposed the model, it becomes necessary to determine if the fitted model is statistically significant. In this case we would like to know whether at least one of the coefficients $\beta_i, 1 \leq i \leq m$, is nonzero which implies that at least one of the variables $x_i, 1 \leq i \leq m$, is useful in explaining the observed variation of the observations $y_i, 1 \leq i \leq n$.

One way of doing this is to test if each individual coefficient $\beta_i, 1 \leq i \leq m$, differs from zero; say through some generalization of the t -test used for simple linear regression. However, if the number of variables is large, the probability of accepting at least one β_i as nonzero is large, even if the significance level of each of the tests is small. Thus, the

overall Type I error of this procedure may be unacceptably large unless the significance level of each individual test is very small.

Then we have the risk of accepting coefficients which are nonzero as zero, and thus possibly omitting significant variables from the model.

In addition, in multiple regression, the multicollinearity problem presents us with additional difficulties since one of the effects of this is to produce large standard errors for the estimated coefficients. This imprecision in estimating individual coefficients may result in accepting all of the coefficients as zero, even if the overall fit is statistically significant. This will be illustrated numerically in Example 5.27.

To overcome the problem of performing many individual tests it would be helpful to have a single statistic to test the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0, \quad (5.175)$$

against

$$H_1 : \text{At least one } \beta_i \neq 0, \quad 1 \leq i \leq m, \quad (5.176)$$

and the ANOVA approach to testing developed for the simple linear regression model suggests that the F -ratio

$$F = \frac{SSR/m}{s^2} \quad (5.177)$$

might be appropriate. In Section 5.8 we shall see that (5.177) is a particular case of a general methodology for testing hypotheses about the general linear model. For now, we will argue for (5.177) on an ad-hoc basis.

We begin by showing that

$$E\left(\frac{SSR}{m}\right) = \sigma^2 + \frac{\langle \mathbf{X}_c \boldsymbol{\beta}_s, \mathbf{X}_c \boldsymbol{\beta}_s \rangle}{m} \quad (5.178)$$

where \mathbf{X}_c is the $n \times m$ matrix of centered regression variables introduced in Example 5.9 and $\boldsymbol{\beta}_s = (\beta_1, \beta_2, \dots, \beta_m)^T$. Since the proof of (5.178) is somewhat involved, the reader may wish to skip the details and proceed to the material after Theorem 5.9 where (5.178) is used to provide a justification for using F to test for the overall significance of the regression.

Theorem 5.8 *Let $\bar{x}_j = \sum_{i=1}^n x_{ij}/n$ be the mean value of the j -th column of \mathbf{X} , and let $\bar{\mathbf{X}}$ denote the matrix whose j -th column is $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)^T$. Then in the general linear model with uncorrelated homoscedastic errors, with variance σ^2*

$$\begin{aligned} E\left(\frac{SSR}{m}\right) &= \sigma^2 + \frac{\langle (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta}, (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta} \rangle}{m} \\ &= \sigma^2 + \frac{\langle \mathbf{X}_c \boldsymbol{\beta}_s, \mathbf{X}_c \boldsymbol{\beta}_s \rangle}{m}. \end{aligned} \quad (5.179)$$

Proof. We begin by noting that

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^m x_{ij} \hat{\beta}_j, \quad (5.180)$$

and from $\partial SSE/\partial\beta_0 = 0$, that

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^m \bar{x}_j \hat{\beta}_j. \quad (5.181)$$

Thus,

$$\hat{y}_i - \bar{y} = \sum_{j=1}^m (x_{ij} - \bar{x}_j) \hat{\beta}_j, \quad 1 \leq i \leq n. \quad (5.182)$$

Using the definition of $\bar{\mathbf{X}}$ and noting that the 0-th column of $\mathbf{X} - \bar{\mathbf{X}}$ is $\mathbf{0}$, (5.182) can be written in vector-matrix form as

$$\hat{\mathbf{Y}} - \bar{\mathbf{Y}} = (\mathbf{X} - \bar{\mathbf{X}}) \hat{\boldsymbol{\beta}} \quad (5.183)$$

where $\bar{\mathbf{Y}} = (\bar{y}, \bar{y}, \dots, \bar{y})^T$. Thus,

$$\begin{aligned} SSR &= \langle \hat{\mathbf{Y}} - \bar{\mathbf{Y}}, \hat{\mathbf{Y}} - \bar{\mathbf{Y}} \rangle = \langle (\mathbf{X} - \bar{\mathbf{X}}) \hat{\boldsymbol{\beta}}, (\mathbf{X} - \bar{\mathbf{X}}) \hat{\boldsymbol{\beta}} \rangle \\ &= \langle \hat{\boldsymbol{\beta}}, (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) \hat{\boldsymbol{\beta}} \rangle = \langle \hat{\boldsymbol{\beta}}, \mathbf{A} \hat{\boldsymbol{\beta}} \rangle \end{aligned} \quad (5.184)$$

where $\mathbf{A} = (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})$.

Now from (5.184) and Theorem 4.14

$$E(SSR) = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \langle \boldsymbol{\beta}, \mathbf{A}\boldsymbol{\beta} \rangle \quad (5.185)$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. We now proceed to simplify \mathbf{A} . First, observe that

$$\bar{\mathbf{X}} = \frac{\mathbf{E}\mathbf{X}}{n} \quad (5.186)$$

where

$$\mathbf{E} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad (5.187)$$

is an $n \times n$ matrix consisting of all ones. Thus,

$$\mathbf{X} - \bar{\mathbf{X}} = \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{X} \quad (5.188)$$

and $(\mathbf{X} - \bar{\mathbf{X}})^T = \mathbf{X}^T (\mathbf{I}_n - \mathbf{E}/n)$ since \mathbf{E} is symmetric. Using the fact that $\mathbf{E}^2/n^2 = \mathbf{E}/n$, we get that

$$(\mathbf{I}_n - \mathbf{E}/n)^2 = \mathbf{I}_n - \frac{2\mathbf{E}}{n} + \frac{\mathbf{E}^2}{n} = \mathbf{I}_n - \frac{\mathbf{E}}{n} \quad (5.189)$$

giving

$$\mathbf{A} = \mathbf{X}^T \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{X}. \quad (5.190)$$

Also,

$$\mathbf{A}\Sigma = \sigma^2 \left[\mathbf{X}^T \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{X} \right] (\mathbf{X}\mathbf{X}^T)^{-1} \quad (5.191)$$

and using (4.32)

$$\begin{aligned} \text{tr}(\mathbf{A}\Sigma) &= \text{tr}(\Sigma\mathbf{A}) = \sigma^2 \text{tr} \left[\mathbf{X} (\mathbf{X}^T \mathbf{X}^{-1}) \mathbf{X}^T \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \right] \\ &= \sigma^2 \text{tr} \left[\mathbf{H} - \frac{\mathbf{H}\mathbf{E}}{n} \right] \\ &= \sigma^2 \left[\text{tr}(\mathbf{H}) - \text{tr} \left(\frac{\mathbf{H}\mathbf{E}}{n} \right) \right]. \end{aligned} \quad (5.192)$$

Since $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(1, 1, \dots, 1)^T \equiv \mathbf{1}$ is the 0-th column of \mathbf{X} , $\mathbf{H}\mathbf{1} = \mathbf{1}$, so that $\mathbf{H}\mathbf{E} = \mathbf{E}$. Thus,

$$\begin{aligned} \text{tr}(\mathbf{H}) - \text{tr} \left(\frac{\mathbf{H}\mathbf{E}}{n} \right) &= \text{tr}(\mathbf{H}) - \text{tr} \left(\frac{\mathbf{E}}{n} \right) \\ &= m + 1 - 1 = m \end{aligned} \quad (5.193)$$

so that $\text{tr}(\mathbf{A}\Sigma) = m\sigma^2$. Finally, using this in (5.185) gives

$$E \left(\frac{SSR}{m} \right) = \sigma^2 + \frac{\langle (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta}, (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta} \rangle}{m}. \quad (5.194)$$

Now, $(\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta} = \mathbf{X}_c \boldsymbol{\beta}_s$ because the first column of $\mathbf{X} - \bar{\mathbf{X}}$ is zero. Thus,

$$\langle (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta}, (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta} \rangle = \langle \mathbf{X}_c \boldsymbol{\beta}_s, \mathbf{X}_c \boldsymbol{\beta}_s \rangle \quad (5.195)$$

so that

$$E \left(\frac{SSR}{m} \right) = \sigma^2 + \frac{\langle \mathbf{X}_c \boldsymbol{\beta}_s, \mathbf{X}_c \boldsymbol{\beta}_s \rangle}{m}. \quad (5.196)$$

as required. ■

From (5.196) we see that if $\boldsymbol{\beta}_s = 0$ then $E(SSR/m) = E(s^2) = \sigma^2$. Hence, if $\boldsymbol{\beta}_s \neq 0$, then on average $SSR/m > \sigma^2$ since $E(SSR/m) > \sigma^2$. Thus, large values of the F -ratio $(SSR/m)/s^2$ suggest that we should reject the hypothesis H_0 that $\beta_1 = \beta_2 = \dots = \beta_m = 0$.

To determine the critical values for rejecting H_0 we need to determine the distribution of F . In order to do this we will need to make our standard assumptions that the errors are independent $N(0, \sigma^2)$ random variables.

Theorem 5.9 *If the errors in (5.1) are independent $N(0, \sigma^2)$, and if $\boldsymbol{\beta}_s = 0$, then F has an F -distribution with $(m, n - m - 1)$ degrees of freedom.*

Proof. Note from (5.184) and (5.190) that

$$SSR = \left\langle \hat{\mathbf{Y}}, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \hat{\mathbf{Y}} \right\rangle. \quad (5.197)$$

To simplify (5.197) we note that

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \mathbf{H}(\mathbf{1}\beta_0 + \mathbf{X}_v\beta_s + \boldsymbol{\varepsilon}) = \beta_0\mathbf{H}\mathbf{1} + \mathbf{H}\mathbf{X}_v\beta_s + \mathbf{H}\boldsymbol{\varepsilon}. \quad (5.198)$$

Since $\beta_s = \mathbf{0}$, $\mathbf{H}\mathbf{X}_v\beta_s = \mathbf{0}$ and $\mathbf{H}\mathbf{1} = \mathbf{1}$, $\mathbf{H}\mathbf{Y} = \beta_0\mathbf{H}\mathbf{1} + \mathbf{H}\boldsymbol{\varepsilon}$.

Thus,

$$\left\langle \hat{\mathbf{Y}}, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \hat{\mathbf{Y}} \right\rangle = \left\langle \beta_0\mathbf{1} + \mathbf{H}\boldsymbol{\varepsilon}, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) (\beta_0\mathbf{1} + \mathbf{H}\boldsymbol{\varepsilon}) \right\rangle \quad (5.199)$$

Since $(\mathbf{I}_n - \mathbf{E}/n)\mathbf{1} = \mathbf{0}$ (5.199) simplifies to

$$\left\langle \beta_0\mathbf{1} + \mathbf{H}\boldsymbol{\varepsilon}, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{H}\boldsymbol{\varepsilon} \right\rangle. \quad (5.200)$$

Using the fact that $\mathbf{I}_n - \mathbf{E}/n$ is symmetric and $(\mathbf{I}_n - \mathbf{E}/n)\mathbf{I}_n = \mathbf{0}$ in (5.200)

$$SSR = \left\langle \mathbf{H}\boldsymbol{\varepsilon}, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{H}\boldsymbol{\varepsilon} \right\rangle = \left\langle \boldsymbol{\varepsilon}, \mathbf{H} \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{H}\boldsymbol{\varepsilon} \right\rangle. \quad (5.201)$$

Due to the symmetry and idempotency of \mathbf{H} (i.e., $\mathbf{H}^2 = \mathbf{H}$), and the fact that $\mathbf{H}\mathbf{E} = \mathbf{E}\mathbf{H} = \mathbf{E}$ (this follows from $\mathbf{H}\mathbf{1} = \mathbf{1}$) (5.201) becomes

$$SSR = \langle \boldsymbol{\varepsilon}, \mathbf{B}\boldsymbol{\varepsilon} \rangle \quad (5.202)$$

where $\mathbf{B} = \mathbf{H} - \mathbf{E}/n$.

Now

$$\mathbf{B}^T = \mathbf{H}^T - \frac{\mathbf{E}^T}{n} = \mathbf{H} - \frac{\mathbf{E}}{n} \quad (5.203)$$

and

$$\begin{aligned} \mathbf{B}^2 &= \left(\mathbf{H} - \frac{\mathbf{E}}{n} \right) \left(\mathbf{H} - \frac{\mathbf{E}}{n} \right) = \mathbf{H}^2 - \frac{\mathbf{H}\mathbf{E}}{n} - \frac{\mathbf{E}\mathbf{H}}{n} + \frac{\mathbf{E}^2}{n^2} \\ &= \mathbf{H} - \frac{2\mathbf{H}\mathbf{E}}{n} + \frac{\mathbf{E}}{n} = \mathbf{H} - \frac{2\mathbf{E}}{n} + \frac{\mathbf{E}}{n} = \mathbf{H} - \frac{\mathbf{E}}{n} = \mathbf{B}. \end{aligned} \quad (5.204)$$

Thus, \mathbf{B} is symmetric and idempotent so that $\text{rank}(\mathbf{H} - \mathbf{E}/n) = \text{tr}(\mathbf{H} - \mathbf{E}/n) = m + 1 - 1 = m$. Hence, using the spectral theorem

$$\mathbf{B} = \mathbf{H} - \frac{\mathbf{E}}{n} = \mathbf{Q}^T \left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \mathbf{Q} \quad (5.205)$$

so that

$$SSR = \left\langle \mathbf{Q}\boldsymbol{\varepsilon}, \left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \mathbf{Q}\boldsymbol{\varepsilon} \right\rangle = \left\langle \mathbf{Z}, \left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \mathbf{Z} \right\rangle \quad (5.206)$$

where

$$\mathbf{Z} = \mathbf{Q}\boldsymbol{\varepsilon}. \quad (5.207)$$

Thus,

$$SSR = \sum_{i=1}^m Z_i^2 \quad (5.208)$$

where $Z_i \sim N(0, \sigma^2)$. (This follows from (5.207) because $\epsilon \sim N(\mathbf{0}, \mathbf{I}_n)$ and \mathbf{Q} is orthogonal.) Thus, $Z_i/\sigma \sim N(0, 1)$ and $SSR/\sigma^2 \sim \chi^2(m)$.

To complete the proof we must show that SSR and s^2 are independent random variables. If this is the case, then

$$F = \frac{SSR/m}{s^2} = \frac{SSR/m\sigma^2}{s^2/\sigma^2} = \frac{\chi^2(m)/m}{\chi^2(n-m-1)/(n-m-1)} \quad (5.209)$$

is the ratio of two independent mean squares, hence it has an F -distribution.

To show the independence of SSR and s^2 it suffices to prove that SSR is independent of the residuals $\hat{\epsilon}_i, 1 \leq i \leq n$. Now from (5.202)

$$\begin{aligned} SSR &= \left\langle \mathbf{H}\epsilon, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{H}\epsilon \right\rangle \\ &= \left\langle \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{H}\epsilon, \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \mathbf{H}\epsilon \right\rangle. \end{aligned} \quad (5.210)$$

Letting

$$\mathbf{W} = \mathbf{H} \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) \epsilon \equiv \mathbf{A}\epsilon \quad (5.211)$$

and

$$\hat{\epsilon} = (\mathbf{I}_n - \mathbf{H}) \epsilon \equiv \mathbf{B}\epsilon, \quad (5.212)$$

it suffices to show that \mathbf{W} and ϵ are independent random vectors. For this, letting

$$\mathbf{V} = \begin{pmatrix} \mathbf{W} \\ \hat{\epsilon} \end{pmatrix}^T \text{ and } \mathbf{V} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \epsilon \quad (5.213)$$

and arguing as in Theorem 5.5 it now suffices to show that $\mathbf{AB}^T = \mathbf{0}$. Now, since $\mathbf{I}_n - \mathbf{H}$ is symmetric

$$\mathbf{AB}^T = \mathbf{H} \left(\mathbf{I}_n - \frac{\mathbf{E}}{n} \right) (\mathbf{I}_n - \mathbf{H}) = \mathbf{H} \left(\mathbf{I}_n - \mathbf{H} - \frac{\mathbf{E}}{n} + \frac{\mathbf{EH}}{n} \right). \quad (5.214)$$

Since $\mathbf{EH} = \mathbf{E}$ we get

$$\mathbf{AB}^T = \mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \mathbf{0}. \quad (5.215)$$

■

Using Theorem 5.9 we see that if $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ H_0 is rejected at level α if

$$F > f_{\alpha, m, n-m-1}. \quad (5.216)$$

This is the classical F -test for the overall significance of the regression. Although the F -test given by (5.216) can be strictly used only if the errors are independent $N(0, \sigma^2)$ in the GLM, it is generally used even if this is not known to be true.

As for simple linear regression, this can frequently be justified, at least for large sample sizes, by central limit theorem considerations. The F -test developed above leads

us to introduce the following ANOVA table as a way of summarizing the appropriate statistics.

Table 5.8 ANOVA Table for the General Linear Model

Source	df	Sum of Squares	Mean Squares	F
Regression	m	SSR	$MSR = \frac{SSR}{m}$	$\frac{MSR}{s^2}$
Residual	$n - (m + 1)$	SSE	$MSE = \frac{SSE}{n - m - 1}$	
Total	$n - 1$	SST		
		R^2	\bar{R}^2	

5.6.2 The Coefficient of Multiple of Determination

As in the case of simple linear regression we can think of the F -test, as testing for the significance of the coefficient of multiple determination defined by

$$R^2 \equiv \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (5.217)$$

because

$$\begin{aligned}
 F &= \frac{SSR}{SSE} \left(\frac{n - m - 1}{m} \right) \\
 &= \left[\left(\frac{SSR}{SST} \right) / \left(\frac{SSE}{SST} \right) \right] \left(\frac{n - m - 1}{m} \right) \\
 &= \left[\frac{R^2}{1 - R^2} \right] \left(\frac{n - m - 1}{m} \right)
 \end{aligned} \quad (5.218)$$

which is a monotone increasing function of R^2 .

Since it follows from the decomposition of SST given in Eq. (5.141) that $0 \leq R^2 \leq 1$, as for simple linear regression, large values of R^2 close to one generally indicate good fits, while “small” values suggests poor ones. Because of this, most practitioners of regression analysis routinely examine R^2 as an important output statistic. The reader is cautioned however, that R^2 can be made arbitrarily close to one merely by adding variables to the regression model (5.1), even if the variables are statistically insignificant. In fact, if we have n variables (linearly independent) and n observations we get a perfect fit. This problem of *overfitting* can generally be detected, because the F -ratio may then be insignificant (say at $\alpha = 0.05$) or only marginally significant. To see why this is so, consider SSE/SST in (5.218). Now SST is fixed by the data, but SSE can be decreased by adding variables since we then minimize $\langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle$ over a larger set of β 's. Thus SSE/SST is a nondecreasing function of the number of variables, so R^2 is nondecreasing. Because of this possibility of artificially inflating R^2 one can modify R^2 in (5.217) by replacing SSE and SST by their mean squares $SSE/(n - m - 1)$ and $SST/(n - 1)$ respectively giving the *adjusted* R^2 ,

$$\bar{R}^2 = 1 - \frac{SSE/(n - m - 1)}{SST/(n - 1)}. \quad (5.219)$$

Note that because of the factor $n - m - 1$ that the addition of a variable does not necessarily increase \bar{R}^2 since it decreases as m increases to compensate for the decrease in SSE .

As for simple linear regression we complete the ANOVA table, Table 5.8 by adding a find row listing the values of R^2 and \bar{R}^2 .

Example 5.22 (ANOVA tables for Examples 5.12-5.17) Here we present the ANOVA tables for Examples 5.12-5.17 which will allow us to make further inferences concerning the models proposed there.

- (1) Housing data: For the housing price data in Example 5.12 we find that the ANOVA table is:

Table 5.9 ANOVA for Housing Data

Source	df	Sum of Squares	Mean Squares	F
Regression	2	1.67531×10^{10}	8.3766×10^9	138.99
Residual	12	7.23185×10^8	6.0265×10^7	
Total	14	1.74763×10^{10}		
		$R^2 = 0.959$	$\bar{R}^2 = 0.952$	

From Table 5.9, $F = 138.99$ which again is significant at $< 0.1\%$ level so we accept the hypotheses that at least one of the variables, square footage or age is significant in explaining the observed variation in housing prices.

- (2) Birthweight data: The ANOVA table for the model given by (5.103) is:

Table 5.10 ANOVA for Birth Weight Data

Source	df	Sum of Squares	Mean Squares	F
Regression	3	1,217,562	405,854	13.26
Residual	20	612,311	30,616	
Total	23	1,829,873		
		$R^2 = 0.66$	$\bar{R}^2 = 0.615$	

Again F is significant at $< 0.1\%$ level which leads us to accept the hypothesis that at least one of x_1, x_2 or x_3 is significant in explaining the variation in birth weights. From our discussion in Chapter 3 we certainly expect x_1 (gestation period) to be significant but as yet we cannot infer anything about the effect of sex. From Table 5.6 we see that the VIFs for $\hat{\beta}_2$ and $\hat{\beta}_3$ are quite large, suggesting a strong multicollinearity between these factors. As will be shown, this problem confounds the interpretation of the significance of x_2 and x_3 .

- (3) Hald data: The ANOVA table for the Hald data is:

Table 5.11 ANOVA for Hald Data

Source	df	Sum of Squares	Mean Squares	F
Regression	4	2,667.90	666.97	111.48
Residual	8	47.86	5.98	
Total	12	2,715.76		
		$R^2 = 0.982$	$\bar{R}^2 = 0.974$	

As in our previous examples, F is significant at $< 0.1\%$ level and so we accept the hypothesis that at least one of x_1 - x_4 is significant in explaining the variation in

heat evolved. The large values of R^2 and \bar{R}^2 indicate that essentially all of the variation in y can be explained by the model (5.104). However, the large VIFs for $\hat{\beta}_2$ and $\hat{\beta}_4$ suggests a possible problem in interpreting the estimated coefficients $\hat{\beta}_2$ and $\hat{\beta}_4$.

(4) Longley data:

Table 5.12 ANOVA for Longley Data

Source	df	Sum of Squares	Mean Squares	F
Regression	6	184,172,402	30,695,400	330.29
Residual	9	836,424	92,936	
Total	15	185,008,826		
		$R^2 = 0.995$	$\bar{R}^2 = 0.992$	

From Table 5.12 we see that $F = 330.29$ which is significant at $< 0.1\%$ level indicating that at least one of the variables x_1 - x_6 is significant in explaining the observed data. Also, the large values of R^2 and \bar{R}^2 indicates that there is almost a perfect fit using the predictors x_1 - x_6 . The possibility that fewer variables may be sufficient will be discussed in the next section.

As for simple linear regression R^2 has a correlation interpretation. It is the square of the sample correlation between $\hat{\mathbf{y}}$ and \mathbf{y} .

5.7 Confidence Intervals and t -Tests for the Coefficients

5.7.1 Confidence Intervals

If the overall regression is found to be significant, attention then generally turns to a consideration of which of the individual coefficients is contributing to the fit. This analysis may be carried out using tests developed from confidence intervals for the regression coefficients as was done for simple linear regression in Chapter 3. The confidence intervals we develop are strictly valid only if the standard normality conditions are met. However, under appropriate conditions on the errors, they may be valid asymptotically for large sample sizes n , and they are routinely used in practice for screening the contribution of individual variables even when little is known about the errors.

To derive these intervals let δ_j denote the j -th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Then, as we showed in Theorem 5.5

$$T = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{\delta_j}} \quad (5.220)$$

has a t -distribution with $n-m-1$ degrees of freedom. Then using standard manipulations as in Chapters 2 and 3 we find that a $(1-\alpha) \times 100\%$ confidence interval for β_j is given by

$$\left(\hat{\beta}_j - t_{n-m-1, \alpha/2} s\sqrt{\delta_j}, \hat{\beta}_j + t_{n-m-1, \alpha/2} s\sqrt{\delta_j} \right). \quad (5.221)$$

If $m = 1$, these intervals are easily shown to be those derived in Chapter 3 for the coefficients of the simple linear regression model.

5.7.2 *t*-Tests

Using the confidence intervals given in (5.221) we can develop a test of

$$H_0 : \beta_j = b_j \quad (5.222)$$

against

$$H_1 : \beta_j \neq b_j. \quad (5.223)$$

The critical region for a level α test is found by rejecting H_0 if b_j is not in the confidence interval (5.221). Simple calculations show that this region is given by

$$\frac{\hat{\beta}_j - b_j}{s\sqrt{\delta_j}} < -t_{n-m-1, \alpha/2} \quad \text{or} \quad \frac{\hat{\beta}_j - b_j}{s\sqrt{\delta_j}} > t_{n-m-1, \alpha/2} \quad (5.224)$$

or equivalently

$$|T_j| > t_{n-m-1, \alpha/2} \quad (5.225)$$

where $T_j = (\hat{\beta}_j - b_j) / s\sqrt{\delta_j}$.

When $b_j = 0$, we have a test for the significance of an individual regression coefficient. That is, we reject $H_0 : \beta_j = 0$ at level α if

$$\left| \frac{\hat{\beta}_j}{s\sqrt{\delta_j}} \right| > t_{n-m-1, \alpha/2}. \quad (5.226)$$

If $n-m-1 > 20$, one can use an “eyeball” test for testing the significance of $\hat{\beta}_j$. Because, for large n , T_j is approximately $N(0, 1)$, a 5% level test is to reject H_0 if

$$|T_j| > t_{n-m-1, 0.025} \simeq z_{0.025} = 1.96 \simeq 2. \quad (5.227)$$

This allows one to quickly screen computer output for the significance of regression coefficients. If no significant model violations or multicollinearity is present, then one might consider removing any variable for which $|T_j| < 2$. In particular, when serious multicollinearity is present, then the regression may be significant as measured by the *F*-test, but all or almost all of the variables may appear to be insignificant as measured by *t*-tests. Conversely, as shown by Freedman in [37], when a model has a large number of possible variables, then many variables may appear to be significant, even if they are random noise! When using *t*-tests proceed with caution!

By using one-sided confidence intervals, one can develop tests of $H_0 : \beta_j = b_j$ against the one-sided alternatives

$$H_1 : \beta_j > b_j \quad \text{or} \quad H_1 : \beta_j < b_j. \quad (5.228)$$

In the first case the critical region for a level α test is to reject H_0 if

$$T_j > t_{n-m-1, \alpha} \quad (5.229)$$

while in the second it is to reject H_0 if

$$T_j < t_{n-m-1, \alpha}. \quad (5.230)$$

Several examples illustrating these ideas are given next.

Example 5.23 (Format for listing t -values) Before presenting numerical results for Examples 5.24-5.27 we give a standard format for presenting t values for each estimated regression coefficient. This format, or slight modifications of it, is standard output from most widely available statistical packages.

Table 5.13 t -values for Estimated Regression Coefficients				
Predictor	Coefficient	S.E.	t -statistic	p -value
constant	$\hat{\beta}_0$	$\hat{\sigma} \left(\hat{\beta}_0 \right)$	$t_0 = \hat{\beta}_0 / \hat{\sigma} \left(\hat{\beta}_0 \right)$	$P \{T_0 > t_0 \}$
x_1	$\hat{\beta}_1$	$\hat{\sigma} \left(\hat{\beta}_1 \right)$	$t_1 = \hat{\beta}_1 / \hat{\sigma} \left(\hat{\beta}_1 \right)$	$P \{T_1 > t_1 \}$
x_2	$\hat{\beta}_2$	$\hat{\sigma} \left(\hat{\beta}_2 \right)$	$t_2 = \hat{\beta}_2 / \hat{\sigma} \left(\hat{\beta}_2 \right)$	$P \{T_2 > t_2 \}$
.
.
.
x_m	$\hat{\beta}_m$	$\hat{\sigma} \left(\hat{\beta}_m \right)$	$t_m = \hat{\beta}_m / \hat{\sigma} \left(\hat{\beta}_m \right)$	$P \{T_m > t_m \}$

The first column lists the variables in the model. The second column gives the least squares estimates $\hat{\beta}_i$ of $\beta_i, 0 \leq i \leq m$, while the third column gives their standard errors. The t values are given in the fourth column and their significance levels (to three significant figures) are given in the last. Unless otherwise stated, we will assume a significance level of 5%, so any value $p < 0.05$ indicates a coefficient significantly different from zero. Sometimes an additional column listing VIFs is also given.

Example 5.24 (t values for Longley data) Using the format in Example 5.23 the t values and their significance levels are given below.

Table 5.14 t values for Longley data					
Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value	VIF
constant	-3,482,259	890,420	-3.91	*0.004	
x_1	15.06	84.91	0.18	0.863	135.5
x_2	-0.03582	0.03349	-1.07	0.313	1788.5
x_3	-2.0202	0.4884	-4.14	*0.003	33.6
x_4	-1.0332	0.2143	-4.82	*0.001	3.6
x_5	-0.0511	0.2261	-0.23	0.826	399.2
x_6	1829.2	455.5	4.02	*0.003	759.0

From these results we see that $\beta_0, \beta_3, \beta_4$ and β_6 appear to be significantly different from zero at the 5% level (indicated by *), while β_1, β_2 and β_5 appear to be zero. However, the apparent multicollinearity in the data and our comments concerning multiple t -tests suggests caution in eliminating x_1, x_2 and x_5 from the model, since doing this one can bias the remaining coefficients. Notwithstanding those caveats we refit the data by regressing y on the variables x_3, x_4 and x_6 . The ANOVA table and t statistics are given below.

Table 5.15 ANOVA table for reduced Longley data

Source	df	Sum of Squares	Mean Squares	<i>F</i>
Regression	3	183,685,465	61,228,488	555.21
Residual	12	1,323,361	110,280	
Total	15	185,008,826		
		$R^2 = 0.993$	$\bar{R}^2 = 0.991$	

Table 5.16 *t* statistics for reduced Longley data

Predictor	Coefficient	S.E. Coeff.	<i>t</i> -statistic	<i>p</i> -value	VIF
constant	−1,797,221	68,642	−26.18	0.000	
<i>x</i> ₃	−1.4697	0.1671	−8.79	0.000	3.3
<i>x</i> ₄	−0.7723	0.1837	−4.20	0.001	2.2
<i>x</i> ₆	956.38	35.52	26.92	0.000	3.9

From Table 5.15 we see that the ANOVA tables are quite similar with virtually the same values of R^2 and \bar{R}^2 . Hence, the reduced model appears to explain the observed data as well as the full model. In addition, the standard errors of the coefficients are significantly smaller than in the full model giving highly significant *t* values. Moreover, there has been a substantial reduction in the VIFs indicating that the multicollinearity problem has been substantially alleviated. It appears that in eliminating the variables x_1, x_2 and x_5 from the full model we have obtained a model which seems to be much more reliable than the original one.

It is interesting to ask if this is the “best” model. This is a topic we will take up in more detail in Chapter 8. For comparison purposes we used a standard *stepwise regression variable selection* procedure on the full Longley data. As we discuss later, such methods purport to select the best set of predictors from a given proposed set. Using this procedure only the variables x_2 and x_3 were selected putting in a variable deemed insignificant in our original analysis and omitting two variables that were highly significant in the second. Further details are given in Tables 5.17-5.18.

Table 5.17 ANOVA Table for “best” Longley model

Source	df	Sum of Squares	Mean Squares	<i>F</i>
Regression	2	181,429,761	90,714,881	329.50
Residual	13	3,579,065	275,313	
Total	15	185,008,826		
		$R^2 = 0.981$	$\bar{R}^2 = 0.978$	

Table 5.18 *t* statistics for “best” Longley model

Predictor	Coefficient	S.E. Coeff.	<i>t</i> -statistic	<i>p</i> -value	VIF
constant	52,382.20	573.5	91.33	0.000	
<i>x</i> ₂	0.037840	0.001711	22.12	0.000	1.6
<i>x</i> ₃	−0.5436	0.1820	−2.99	0.010	1.6

Again we observe that the ANOVA table is similar to Tables 5.9 and 5.15 and the overall fit appears to be about the same as the full and reduced models. But there is a substantial change in the intercept and a change in sign of the coefficient of x_2 from negative in the full model to positive in the “best” model.

Since x_2 is the GNP, intuitively we expect total employment to increase as GNP increases and the “best” model shows this while the full model gives a result which is counterintuitive. Again, this is another consequence of multicollinearity. Clearly, further tools are needed to assess the appropriateness of the proposed models and these will be developed as we proceed.

Example 5.25 (*t* statistics for drink delivery data) Continuing our analysis of the drink delivery data we give the *t* statistics in Table 5.19.

Table 5.19 *t* statistics for drink delivery data

Predictor	Coefficient	S.E. Coeff.	<i>t</i> -statistic	<i>p</i> -value	VIF
constant	2.341	1.097	2.13	0.044	
x_1	1.6159	0.1707	9.46	0.000	3.1
x_2	0.014385	0.003613	3.98	0.001	3.1

From Table 5.19 we see that both β_1 and β_2 are significantly different from zero so that including “distance walked” appears to be useful in predicting the time it takes to deliver drinks in addition to the number of cases. Without further examination, we can conclude that the model

$$Y = 2.341 + 1.6159x_1 + 0.014385x_2 + \varepsilon$$

appears to be quite acceptable for explaining the variation in delivery times.

Example 5.26 (*t* statistics for birth weight data) In example 5.15 we found that the overall model of two lines was highly significant in explaining the observed variation in birth weights. From our analysis in Chapter 3 we expect gestation period to be a significant predictor but is “sex” as well? To evaluate this we give the *t*-statistics in Table 5.20.

Table 5.20 *t* statistics for birth weight data

Predictor	Coefficient	S.E. Coeff.	<i>t</i> -statistic	<i>p</i> -value	VIF
constant	-2,142	1,189	-1.80	0.087	
x_1	130.40	30.66	4.25	0.000	2.1
x_2	1,042	1647	0.63	0.534	477.6
x_3	-22.73	42.67	-0.53	0.600	473.3

From Table 5.20 we see that the results generally confirm the model using “gestation period” as the only predictor. β_1 is significantly different from zero, while β_0 is only marginally so. However, the effect of sex is not clear, since the large VIFs for $\hat{\beta}_2$ and $\hat{\beta}_3$ indicate a strong collinearity between these two variables.

Of course, simply eliminating the variables x_2 and x_3 does not seem appropriate, since our previous analysis suggests that the sex of a baby influences its birth weight. Since (5.101) represents two lines, eliminating x_2 would yield a model given by one line so it is not logical to include an interaction term which provides for a difference in slope in identical lines. Hence, if we eliminate one of the variables it should be x_3 first. Doing

this and refitting the data gave the following ANOVA table and t statistics.

Table 5.21 ANOVA table for reduced birth weight data

Source	df	Sum of Squares	Mean Squares	F
Regression	2	1,209,033	604,517	20.45
Residual	21	620,840	29,564	
Total	23	1,829,873		

$R^2 = 0.661$ $\bar{R}^2 = 0.628$

Table 5.22 t statistics for reduced birth weight data

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value	VIF
constant	-1,702.30	747.00	-2.28	0.000	
x_1	119.06	19.23	6.19	0.000	1.0
x_6	182.12	71.09	2.57	0.011	1.0

From Table 5.22 we see that the reduced model is significant in explaining the observed birth weights since $P\{F > 20.45\} < 10^{-3}$. In contrast to the full model, the coefficients β_1 and β_2 are significantly different from zero. Also both VIFs equal one so the effect of multicollinearity seems to have been eliminated and the signs of both coefficients agree with our intuition. As a consequence, the analysis so far indicates that an appropriate model for the birth weight data is two parallel straight lines represented by

$$Y = -1702.3 + 119.06x_1 + 182.12x_2 + \varepsilon.$$

Example 5.27 (Hald data) As discussed in Example 5.16 the Hald data appeared to fit almost perfectly using the predictors x_1 - x_4 . However, the large VIFs suggested possible imprecision in estimating the coefficients.

Table 5.23 t statistics for Hald data

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value	VIF
constant	62.41	70.07	0.89	0.399	
x_1	1.5511	0.7448	2.08	0.071	38.5
x_2	0.5102	0.7238	0.70	0.501	254.4
x_3	0.1019	0.7547	0.14	0.896	46.9
x_4	-0.1441	0.7091	-0.20	0.844	282.5

From Table 5.23 we see that even though the overall fit is highly significant, none of the variables appears to be significant at the 5% level. The only variable which appears marginally significant is x_1 . Since our analysis suggests at least one of x_1 - x_4 should appear in the model, it appears that x_1 is a good choice. In fact, regressing y on x_1 gives the following ANOVA and t statistics.

Table 5.24 ANOVA table for x_1 in Hald data

Source	df	Sum of Squares	Mean Squares	F
Regression	1	1450.1	1450.1	12.60
Residual	11	1265.7	115.1	
Total	12	2715.8		

$R^2 = 0.534$ $\bar{R}^2 = 0.492$

Table 5.25 t statistics for x_1 in Hald data

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value
constant	81.4790	4.9270	16.54	0.000
x_1	1.8687	0.5264	3.55	0.005

From Tables 5.24 and 5.25 we see that the regression is significant at $< 5\%$ and of course β_1 is seen to be significantly different than zero. However, R^2 is only about one-half that of the full model indicating that at least one other variable is important. From Table 5.23 it is reasonable to consider x_2 since it has the second largest t value. As a consequence, we refit the data using x_1 and x_2 and the results are given in Tables 5.26 and 5.27.

Table 5.26 ANOVA table for Hald data using x_1, x_2

Source	df	Sum of Squares	Mean Squares	F
Regression	2	2,657.9	1,328.9	229.5
Residual	10	57.9	5.8	
Total	12	2,715.8		
		$R^2 = 0.979$	$\bar{R}^2 = 0.974$	

Table 5.27 t statistics for Hald data using x_1, x_2

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value	VIF
constant	52.577	2.286	23.00	0.000	
x_1	1.4683	0.1213	12.10	0.000	1.1
x_2	0.66225	0.04585	14.44	0.000	1.1

From Table 5.27 we observe that the model containing x_1 and x_2 is highly significant in explaining the data since $P\{F > 229.5\} < 10^{-3}$ and the R^2 and \bar{R}^2 values are almost the same as in the full model. From Table 5.27 we see that both β_1 and β_2 are significantly different from zero. In addition, the VIFs indicate that the multicollinearity has been alleviated. At present, it appears that the fitted model

$$Y = 52.577 + 1.4688x_1 + 0.66225x_2 + \varepsilon$$

provides a good explanatory model for the Hald data.

From Examples 5.24-5.27 we see that the statistics F , R^2 and t are useful in helping to disentangle the effect of various predictors on a given set of observed data, but they cannot be used in a completely automatic fashion. Despite the fact that modern computers have enabled us to do regressions “at will”, human intervention is still needed to understand possible contradictory results.

As for simple linear regression, additional tools are necessary to further understand complicated relationships. Those will be developed as we proceed.

5.8 The Extra Sum of Squares Principle

5.8.1 The General Linear Hypothesis

So far we have considered two types of tests for the general linear model: the F -test for the overall significance of the regression and t tests for the significance of individual

regression coefficients. In this section we will show that these tests may be considered as particular cases of a test for a *general linear hypothesis* of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad (5.231)$$

against

$$H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b} \quad (5.232)$$

where \mathbf{C} is an $r \times (m+1)$ matrix of rank r . (Thus, $r \leq m+1$.)

The equation $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$ represents r linearly independent constraints

$$\sum_{j=0}^m c_{ij}\beta_j = b_i, \quad 1 \leq i \leq r, \quad (5.233)$$

on the coefficients of the GLM and the purpose of the test we develop is to determine if the data can justify the imposition of these constraints. Rejecting H_0 indicates that the data cannot support the possibility that all r conditions given in (5.233) hold and we would conclude that at least one of them fails.

Before indicating the nature of the test we choose, let us show that by appropriately choosing \mathbf{C} and \mathbf{b} we arrive at the hypotheses for the overall significance of the regression and those for testing the significance of the individual coefficients.

In the first case, letting \mathbf{C} denote the $m \times (m+1)$ matrix

$$\mathbf{C} = \left[\begin{array}{c|cccc} 0 & 1 & 0 & \cdots & 0 \\ \hline 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdot & \cdots & 1 \end{array} \right] \quad (5.234)$$

and $\mathbf{b} = (0, 0, \dots, 0)^T$, then the conditions $\beta_1 = \beta_2 = \dots = \beta_m = 0$ are equivalent to the vector equation

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{0} \quad (5.235)$$

In the second, the hypothesis that $\beta_i = 0$ can be written as

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad (5.236)$$

with $\mathbf{C} = \left[0, \dots, 0, \frac{1}{i\text{-th}}, 0, \dots, 0 \right]$ and $\mathbf{b} = \mathbf{0}$.

As an additional example, suppose that in the model

$$Y_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_{\mathbf{x}} \quad (5.237)$$

we wanted to simultaneously test that $\beta_2 = 0$ and $\beta_3 = \beta_4$, then this could be done by letting

$$\mathbf{C} = \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{array} \right] \quad (5.238)$$

and $\mathbf{b} = (0, 0)^T$ as the reader may easily verify.

As a less arbitrary example, suppose that in the salary model proposed in Example 5.5 we wanted to test the hypothesis that both male and female salary structures were the same. Then this would require testing

$$H_0 : \beta_2 = \beta_3 = 0 \quad (5.239)$$

against

$$H_1 : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0. \quad (5.240)$$

If

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5.241)$$

and $\mathbf{b} = (0, 0)^T$, then these hypotheses can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \quad (5.242)$$

and

$$H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}. \quad (5.243)$$

5.8.2 The F -Test

To develop the test of

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad (5.244)$$

against

$$H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b}, \quad (5.245)$$

we begin by fitting the unconstrained model, where $\mathbf{C}\boldsymbol{\beta} \neq \mathbf{b}$ holds. This is usually called the *full model*. We then fit the model under the condition that H_0 is true, i.e., that $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$. This is usually called the *reduced model*.

Now let SSE_F denote the residual sum of squares from the full model and SSE_R that from the reduced model and note that it will be the case that $SSE_F \leq SSE_R$. However, if H_0 is false, then we would expect the reduction in the residual sum of squares

$$\Delta SSE = SSE_R - SSE_F \quad (5.246)$$

to be significantly greater than the random error σ^2 . Hence, we would expect that a reasonable test would compare ΔSSE to s^2 and reject H_0 if $\Delta SSE/s^2$ was sufficiently large. This test will be based on a generalization of the F -test for the overall significance of the regression based on the F statistic

$$F = \frac{\Delta SSE/r}{s^2}. \quad (5.247)$$

When the errors in the GLM are independent $N(0, \sigma^2)$ random variables and H_0 is true then F has an F -distribution with $(r, n - m - 1)$ degrees of freedom. Thus, the critical region for rejecting H_0 at level α is

$$F > f_{r, n-m-1, \alpha}. \quad (5.248)$$

Before proving this (which leave as optional reading) we will show that this F -test contains as particular cases, the F -test for the overall significance of the regression and

is equivalent to the t -test for the testing the significance of an individual regression coefficient.

Example 5.28 When the standard normality assumptions are met for the GLM, then the F -test given by (5.248) is the same as that given in Section 5.6 for the overall significance of the regression.

To prove this we note that here the reduced model is given by

$$Y_i = \beta_0 + \varepsilon_i, \quad 1 \leq i \leq n, \quad (5.249)$$

so that $SSE_R = \sum_{i=1}^n (y_i - \bar{y})^2 = SST$. Thus,

$$\begin{aligned} \Delta SSE &= SST - SSE_F \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SSR. \end{aligned} \quad (5.250)$$

Hence, the reduction of the residual sum of squares is just the regression sum of squares and F in (5.250) is $(SSR/m)/s^2$ which is the F statistic given previously and so has an F distribution with $(m, n - m - 1)$ degrees of freedom.

To get the result concerning t tests for individual coefficients we will need to use a result which will be established in Theorem 5.10. It will be shown there that

$$\Delta SSE = \left\langle \mathbf{C}\hat{\beta}_F - \mathbf{b}, \mathbf{A} \left(\mathbf{C}\hat{\beta}_F - \mathbf{b} \right) \right\rangle \quad (5.251)$$

where

$$\mathbf{A} = \left[\mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \right]^{-1} \quad (5.252)$$

When testing for the significance of the i -th coefficient β_i , $\mathbf{C} = \left[0, \dots, 0, \underset{i\text{-th}}{1}, 0, \dots, 0 \right]$ and $\mathbf{b} = \mathbf{0}$ so that (show this)

$$\mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T = \delta_i \quad (5.253)$$

where δ_i is the i -th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$. Thus, if $\beta_i = 0$,

$$\Delta SSE = \frac{\hat{\beta}_i^2}{\delta_i} \quad (5.254)$$

and

$$F = \frac{\Delta SSE}{s^2} = \frac{\hat{\beta}_i^2}{s^2 \delta_i} = T_i^2 \quad (5.255)$$

where $T_i = \hat{\beta}_i / s \sqrt{\delta_i}$ is the T statistic for testing $H_0 : \beta_i = 0$. Thus, the F test in (5.255) is equivalent to rejecting H_0 if

$$T_i^2 > f_{1, n-m-1, \alpha} \quad (5.256)$$

and this is equivalent to rejecting H_0 if

$$|T_i| > \sqrt{f_{1, n-m-1, \alpha}} = t_{n-m-1, \alpha/2}. \quad (5.257)$$

For this reason, the t -tests are often referred to as *partial F-tests* [27].

5.8.3 Derivation of the F -Test

Theorem 5.10 *In the GLM with independent $N(0, \sigma^2)$ errors, let SSE_F denote the residual sum of squares from the full model, and let SSE_R denote the residual sum of squares when $H_0: \mathbf{C}\beta = \mathbf{b}$ is true. Then, when H_0 is true*

$$F = \frac{\Delta SSE/r}{s^2} \quad (5.258)$$

has an F -distribution with $(r, n - m - 1)$ degrees of freedom.

In order to prove the Theorem 5.10 we need to derive the formula for ΔSSE given in (5.258). Because of its importance we state this result as a separate lemma.

Lemma 5.2 *If $\hat{\beta}_F$ and $\hat{\beta}_R$ denote the least squares estimates of β in the full and reduced models respectively, then*

(i) $\hat{\beta}_F = \hat{\beta}_R - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C} \hat{\beta}_F - \mathbf{b})$ where

$$\mathbf{A} = [\mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T]^{-1}. \quad (5.259)$$

(ii) $\Delta SSE = SSE_R - SSE_F = \langle \mathbf{C} \hat{\beta}_F - \mathbf{b}, \mathbf{A} (\mathbf{C} \hat{\beta}_F - \mathbf{b}) \rangle$.

Proof. (i) From Theorem 5.1 $\hat{\beta}_F = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, so it remains to find an expression for $\hat{\beta}_R$. We will do this by using the method of Lagrange multipliers. In this case, we want to minimize $g(\beta) = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle$ subject to the r constraints $\mathbf{C}\beta = \mathbf{b}$.

Letting $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_r)$ denote a row vector of *Lagrange multipliers*, $\hat{\beta}_R$ can be found by minimizing the function (*Lagrangean*)

$$L = g(\beta) - 2 \langle \lambda, \mathbf{C}\beta - \mathbf{b} \rangle \quad (5.260)$$

with respect to (β, λ) .

To do this we calculate

$$\frac{\partial L}{\partial \beta} \equiv \left(\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}, \dots, \frac{\partial L}{\partial \beta_m} \right)^T \quad (5.261)$$

and

$$\frac{\partial L}{\partial \lambda} \equiv \left(\frac{\partial L}{\partial \lambda_0}, \frac{\partial L}{\partial \lambda_1}, \dots, \frac{\partial L}{\partial \lambda_m} \right)^T \quad (5.262)$$

and set the resulting partial derivatives to zero.

Now, using (5.260),

$$\begin{aligned} L &= \langle \beta, \mathbf{X}^T \mathbf{X} \beta \rangle - 2 \langle \beta, \mathbf{X}^T \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle - 2 \langle \lambda, \mathbf{C}\beta \rangle - 2 \langle \lambda, \mathbf{b} \rangle \\ &= \langle \beta, \mathbf{X}^T \mathbf{X} \beta \rangle - 2 \langle \beta, \mathbf{X}^T \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle - 2 \langle \mathbf{C}^T \lambda, \beta \rangle - 2 \langle \lambda, \mathbf{b} \rangle \end{aligned} \quad (5.263)$$

so that

$$\frac{\partial L}{\partial \beta} = 2 (\mathbf{X}^T \mathbf{X}) \beta - 2 \mathbf{X}^T \mathbf{y} - 2 \mathbf{C}^T \lambda = 0, \quad (5.264)$$

and

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \mathbf{C}\boldsymbol{\beta} - \mathbf{b} = \mathbf{0}. \quad (5.265)$$

Denoting the solution to (5.264) by $\hat{\boldsymbol{\beta}}_R$ we find that

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \boldsymbol{\lambda} = \boldsymbol{\beta}_F + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \boldsymbol{\lambda}. \quad (5.266)$$

Thus,

$$\mathbf{C}\hat{\boldsymbol{\beta}}_R - \mathbf{b} = \mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b} + \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \boldsymbol{\lambda} \quad (5.267)$$

and using $\mathbf{C}\hat{\boldsymbol{\beta}}_R - \mathbf{b} = 0$ to solve for $\boldsymbol{\lambda}$ gives

$$\boldsymbol{\lambda} = - \left[\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}). \quad (5.268)$$

Substituting this expression for $\boldsymbol{\lambda}$ into (5.266) and using the definition of \mathbf{A} yields

$$\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_F - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}) \quad (5.269)$$

as required.

(ii) For the proof of (ii) we note from (5.269) that for any $\boldsymbol{\beta}$

$$g(\boldsymbol{\beta}) - g(\hat{\boldsymbol{\beta}}_F) = \left\langle \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_F), \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_F) \right\rangle. \quad (5.270)$$

Letting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_R$ in (5.270) shows that (recall $g(\boldsymbol{\beta})$ from (5.20))

$$\begin{aligned} \Delta SSE &= g(\hat{\boldsymbol{\beta}}_R) - g(\hat{\boldsymbol{\beta}}_F) \\ &= \left\langle \mathbf{X}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F), \mathbf{X}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F) \right\rangle \\ &= \left\langle \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}), \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}) \right\rangle \\ &= \left\langle \mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}, \mathbf{A}^T \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}) \right\rangle \\ &= \left\langle \mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}, \mathbf{A}^T \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}) \right\rangle. \end{aligned} \quad (5.271)$$

Now $\mathbf{A} = \mathbf{A}^T$ (verify this) so that

$$\mathbf{A}^T \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} = \mathbf{A} \mathbf{A}^{-1} \mathbf{A} = \mathbf{A} \quad (5.272)$$

which gives

$$\Delta SSE = \left\langle \mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}, \mathbf{A} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}) \right\rangle \quad (5.273)$$

as required. ■

Proof of Theorem 5.10. We begin by showing that $\Delta SSE/\sigma^2$ is $\chi^2(r)$ if H_0 is true. If this is the case, then $\mathbf{Y} \sim \mathbf{N}(\boldsymbol{\beta}_R, \sigma^2 \mathbf{I}_n)$ and $\hat{\boldsymbol{\beta}}_F = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ so that $\hat{\mathbf{r}} = \mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}$ is $\mathbf{N}(E(\hat{\mathbf{r}}), \Sigma(\hat{\mathbf{r}}))$.

But,

$$\begin{aligned} E(\hat{\mathbf{r}}) &= E(\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}) = E[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] - \mathbf{b} \\ &= \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE(\mathbf{Y}) - \mathbf{b} = \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_R - \mathbf{b} \\ &= \mathbf{C}\boldsymbol{\beta}_R - \mathbf{b} = 0. \end{aligned} \quad (5.274)$$

Also,

$$\boldsymbol{\Sigma}(\hat{\mathbf{r}}) = \mathbf{C}\boldsymbol{\Sigma}(\boldsymbol{\beta}_F)\mathbf{C}^T = \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \quad (5.275)$$

so that

$$\mathbf{C}\hat{\boldsymbol{\beta}}_R - \mathbf{b} \sim \mathbf{N}\left(0, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\right). \quad (5.276)$$

Since $\mathbf{A} = [\mathbf{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}]^{-1}$, $\hat{\mathbf{r}} = \mathbf{C}\hat{\boldsymbol{\beta}}_R - \mathbf{b} \sim \mathbf{N}(0, \sigma^2\mathbf{A}^{-1})$ if H_0 is true. Now using the result in (iv) of Theorem 5.5 we find that

$$\frac{\Delta SSE}{\sigma^2} = \frac{\langle \hat{\mathbf{r}}, \mathbf{A}\hat{\mathbf{r}} \rangle}{\sigma^2} \quad (5.277)$$

has a $\chi^2(r)$ distribution.

Since ΔSSE is a function only of $\hat{\boldsymbol{\beta}}_F$, it follows again from (i) of Theorem 5.5 that ΔSSE and s^2 are independent. Thus,

$$F = \frac{\Delta SSE / r\sigma^2}{SSE_F / (n - m - 1)\sigma^2} \quad (5.278)$$

has an F -distribution with $(r, n - m - 1)$ degrees of freedom. ■

Using the basic decomposition of $SST = SSE + SSR$ we also find that

$$\Delta SSE = SSR_F - SSR_R \quad (5.279)$$

so that ΔSSE is also the *increase in the regression sum of squares* due to the failure of H_0 . Thus,

$$SSR_F = SSR_R + \Delta SSE. \quad (5.280)$$

In this case ΔSSE is often referred to as the *extra sum of squares* contributed to the regression sum of squares due to the failure of H_0 . For instance, if $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{m-1}, 0)^T$ (i.e., $\beta_m = 0$) and the true model has $\boldsymbol{\beta} = \boldsymbol{\beta}_F$, then ΔSSE is the *extra regression sum of squares* due to the addition of β_m to the model. For this reason the F -test given by (5.278) is sometimes called the *extra sum of squares principle*. We now present a couple of numerical examples illustrating its use.

Example 5.29 As we have seen in Example 5.12 (housing data), the test for the overall regression is very significant. Now, we consider the question whether there is a significant increase in the variation explained by the additional term $x_3 = (\text{Age})^2 = x_2^2$.

The ANOVA table can be made by the decomposition of the sum of squares as follows:

Table 5.28 ANOVA for Housing Data				
Source	df	Sum of Squares	Mean Squares	F
Regression	3	16,873,260,084	5,624,420,028	102.60
x_1	(1)	15,398,166,924	15,398,166,924	(280.88)
$x_2 x_1$	(1)	1,354,937,318	1,354,937,318	(24.71)
$x_3 x_1, x_2$	(1)	120,155,842	120,155,842	(2.19)
Residual	11	603,029,249	54,820,841	
Total	14	17,476,289,333		
		$R^2 = 0.965$	$\bar{R}^2 = 0.956$	

To test the null hypothesis H_0 : the addition of x_3 (or β_3) to the model does not significantly improve the prediction of Y , we must calculate the extra sum of squares due to the addition of $\beta_3 x_3$ in the model. This sum of squares can be calculated simply from the formula given in Eq (5.279) and also given in Table 5.28,

$$\Delta SSE = SSR_F - SSR_R = 120,155,842.$$

Therefore, we compute the partial F -statistic as

$$F(x_3|x_1, x_2) = \frac{SS(x_3|x_1, x_2)}{MS \text{ Residual}(x_1, x_2, x_3)} = \frac{120,155,842}{54,820,841} \simeq 2.19,$$

and this F -statistic has an F -distribution with 1 and 11 degrees of freedom under H_0 , so we would not reject H_0 at $\alpha = 10\%$.

Example 5.30 (Joint Confidence regions) An interesting consequences of the previous calculations is a method for obtaining joint confidence regions for the regression coefficients $(\beta_0, \beta_1, \dots, \beta_m)^T$. From (5.276) letting $\mathbf{C} = \mathbf{I}_{m+1}$, we see that the quadratic form

$$Q = \langle \hat{\beta} - \beta, (\mathbf{X}^T \mathbf{X})^{-1} (\hat{\beta} - \beta) \rangle \quad (5.281)$$

has a chi-square distribution with $m + 1$ degrees of freedom. Hence,

$$F = \frac{Q/(m+1)}{s^2} = \frac{Q/(m+1)}{SSE/(n-m-1)} \quad (5.282)$$

has an F -distribution with $(m+1, n-m-1)$ degrees of freedom.

Hence,

$$P\{F \leq f_{\alpha, (m+1, n-m-1)}\} = 1 - \alpha \quad (5.283)$$

so that the region $\{F \leq f_{\alpha, (m+1, n-m-1)}\}$ is a $(1 - \alpha)$ 100% joint confidence region for $(\beta_0, \beta_1, \dots, \beta_m)^T$.

Specializing to the case $m = 1$ gives the joint confidence region in (3.121) for the parameters $(\beta_0, \beta_1)^T$ in the SLR model, which is an elliptically shaped region. We leave the details to the reader.

5.9 Prediction

5.9.1 Predicting $E(Y_{\mathbf{x}})$

Once we have fitted the model (5.11) to our data and have decided that the fit is adequate, we can then use the model to make point and interval estimates of both $E(Y_{\mathbf{x}})$ and $Y_{\mathbf{x}}$ as in the case of simple linear regression. We begin with the estimation of $E(Y_{\mathbf{x}})$.

If $\mathbf{x}_0 = (1, x_{0,1}, x_{0,2}, \dots, x_{0,m})$ is a point in the domain of the independent variable \mathbf{x} , then a point estimate of $E(Y_{\mathbf{x}_0})$ is given by

$$\hat{y}_{\mathbf{x}_0} = \hat{\beta}_0 + \sum_{j=1}^m x_{0,j} \hat{\beta}_j = \mathbf{x}_0 \hat{\beta} \quad (5.284)$$

(regarding \mathbf{x}_0 as a row vector and as usual, β as a column vector).

From (5.284) we find that

$$\text{Var}(\hat{Y}_{\mathbf{x}_0}) = \sigma^2 \langle \mathbf{x}_0, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \rangle = \sigma^2 [\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T]. \quad (5.285)$$

And this can be estimated by

$$\hat{\sigma}^2 (\hat{Y}_{\mathbf{x}_0}) = s^2 [\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T] \quad (5.286)$$

where $\hat{\sigma}^2 (\hat{Y}_{\mathbf{x}_0})$ is called the estimated *prediction variance*. In particular, if $\mathbf{x}_0 = \mathbf{x}_i = (1, x_{i,1}, x_{i,2}, \dots, x_{i,m})$ the i -th design point, then

$$\hat{\sigma}^2 (\hat{Y}_{\mathbf{x}_i}) = s^2 [\mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T] = s^2 h_{ii} \quad (5.287)$$

where h_{ii} is the i -th diagonal element of the hat matrix $\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. (Prove this.)

Under the standard normality assumption of the GLM we can use the above results to obtain confidence intervals for $E(Y_{\mathbf{x}_0}) \equiv \mu_{\mathbf{x}_0}$. In fact, since $\hat{Y}_{\mathbf{x}_0}$ is a linear combination of joint multivariate normal random variables, it is normal with

$$E(\hat{Y}_{\mathbf{x}_0}) = \mu_{\mathbf{x}_0} \quad (5.288)$$

and $\text{Var}(\hat{Y}_{\mathbf{x}_0})$ given by (5.286). Thus

$$\frac{\hat{Y}_{\mathbf{x}_0} - \mu_{\mathbf{x}_0}}{\sigma \sqrt{\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T}} \quad (5.289)$$

is $N(0, 1)$ and by the independence of $\hat{\beta}$ and s

$$\frac{\hat{Y}_{\mathbf{x}_0} - \mu_{\mathbf{x}_0}}{s \sqrt{\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T}} \quad (5.290)$$

has a t -distribution with $n - m - 1$ degrees of freedom. From this it follows by standard manipulations that a $(1 - \alpha) \times 100\%$ confidence interval for $\mu_{\mathbf{x}_0}$ is given by

$$\left(\hat{Y}_{\mathbf{x}_0} \pm t_{n-m-1, \alpha/2} s \sqrt{\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T} \right). \quad (5.291)$$

Example 5.31 (Housing data) Suppose that a realtor wishes to construct a 95% confidence interval for the price of a house with $x_1 = 2,400$ square feet and the age of the house $x_2 = 5$ years. Let $\mathbf{x}_0 = (1, 2400, 5)$ and from the previous results

$$\hat{\beta} = \begin{pmatrix} 13239 \\ 60.589 \\ -1726.8 \end{pmatrix}, \quad s = 7763.$$

Using (5.284) the point estimate of the price is, $\hat{y}_{\mathbf{x}_0} = \$150,018.6$. Then, we calculate

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &= \begin{bmatrix} 15 & 31050 & 96 \\ 31050 & 69022500 & 208750 \\ 96 & 208750 & 1090 \end{bmatrix}, \\ (\mathbf{X}^T \mathbf{X})^{-1} &= \begin{bmatrix} 0.977993 & -0.000426 & -0.004463 \\ -0.000426 & 0.000000 & -0.000005 \\ -0.004463 & -0.000005 & 0.002201 \end{bmatrix}. \end{aligned}$$

Hence

$$\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T = 0.0992746.$$

Since $n = 15$ and $m = 3$, $t_{12,0.025} = 2.179$ from Table A.2. Therefore, from (5.291) the 95% confidence interval for the mean predicted value at $x_1 = 2,400$ and $x_2 = 5$ is given by

$$150,018.6 \pm 2.179 (7763) \sqrt{0.09927} = (\$144,688.86, \$155,348.34).$$

5.9.2 Prediction Intervals

If we wish to predict the value of a new observation at \mathbf{x}_0 , then again the point estimate $\hat{Y}_{\mathbf{x}_0}$ is used. If $Y_{\mathbf{x}_0}$ is the true value of $Y_{\mathbf{x}}$ at $\mathbf{x} = \mathbf{x}_0$, then $Y_{\mathbf{x}_0}$ and $\hat{Y}_{\mathbf{x}_0}$ will be independent provided that the observations $Y_{\mathbf{x}_0}, Y_1, Y_2, \dots, Y_n$ are independent, which we assume to be the case. Then,

$$\begin{aligned} \text{Var} (Y_{\mathbf{x}_0} - \hat{Y}_{\mathbf{x}_0}) &= \text{Var} (Y_{\mathbf{x}_0}) + \text{Var} (\hat{Y}_{\mathbf{x}_0}) \\ &= \sigma^2 \left[1 + \mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T \right], \end{aligned} \quad (5.292)$$

so that

$$\frac{Y_{\mathbf{x}_0} - \hat{Y}_{\mathbf{x}_0}}{\sigma \sqrt{1 + \mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T}} \quad (5.293)$$

is a $N(0, 1)$ random variable under the standard normality assumptions of the GLM. From this it follows easily that

$$\frac{Y_{\mathbf{x}_0} - \hat{Y}_{\mathbf{x}_0}}{s \sqrt{1 + \mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T}} \quad (5.294)$$

has a t distribution with $n - m - 1$ degrees of freedom. Using standard manipulations we find that a $(1 - \alpha) \times 100\%$ prediction interval for $Y_{\mathbf{x}_0}$ is given by

$$\left(\hat{Y}_{\mathbf{x}_0} \pm t_{n-m-1, \alpha/2} s \sqrt{1 + \mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T} \right). \quad (5.295)$$

Example 5.32 (Drink delivery data) Suppose that a deliveryman wishes to construct a 90% prediction interval for the delivery time when he has $x_1 = 12$ cases, and distance, $x_2 = 400$ feet. First, in order to find the point estimate of the delivery time, let $\mathbf{x}_0 = (1, 12, 400)$. From the previous results, we have

$$\hat{\beta} = \begin{pmatrix} 2.341 \\ 1.6159 \\ 0.014385 \end{pmatrix}, \quad s = 3.2594$$

so the point estimate, $\hat{Y}_{\mathbf{x}_0} = 27.49$ minutes. Further calculations give

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &= \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}, \\ (\mathbf{X}^T \mathbf{X})^{-1} &= \begin{bmatrix} 0.1132152 & -0.0044486 & -0.00008367 \\ -0.0044486 & 0.0027438 & -0.00004786 \\ -0.00008367 & -0.00004786 & 0.00000123 \end{bmatrix}, \end{aligned}$$

hence we obtain

$$\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T = 0.0717868.$$

Therefore, from (5.295) with $t_{22,0.05} = 1.717$, the 90% prediction interval on the delivery time for $x_1 = 12$ cases and $x_2 = 400$ feet is given by

$$27.49 \pm 1.717 (3.2594) \sqrt{1 + 0.07179} = (21.70, 33.28).$$

5.9.3 Extrapolation

In the case of simple linear regression we showed that the quality of prediction depended essentially on the distance of the data point from the mean of the observation points \bar{x} . Because of this we urged that readers to be cautious in using the fitted model to make predictions outside of the interval $[x_{\min}, x_{\max}]$ where the data were taken. These remarks apply to the case of multiple regression as well; but with some additional problems.

First, it is not immediately clear what the region for extrapolation is. Since the prediction variance is proportional to $\mathbf{x}_0 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T$ points with “large” values of this quantity should be avoided since prediction there will generally be unreliable. In particular, if \mathbf{x}_i is an observed data point, then one can expect that those points for which $h_{ii} = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T$ is largest would lie on the boundary of the set where prediction is reliable. This suggests that points that lie inside the *ellipsoid*

$$\mathbf{x} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \leq \max_{1 \leq i \leq n} h_{ii}, \quad (5.296)$$

may be considered as acceptable places to make predictions, while those outside this region are considered as unacceptable. For example, Figure 5.8 illustrates two points (x_{11}, x_{12}) and (x_{21}, x_{22}) lie within the range of both regressors x_1 and x_2 but outside the joint region of the original data.

In particular, one needs to be careful in using as prediction points those points, whose coordinates are smaller in absolute value than the largest of the absolute values of the components in the data vectors \mathbf{x}_i , since such points will generally not lie in the region

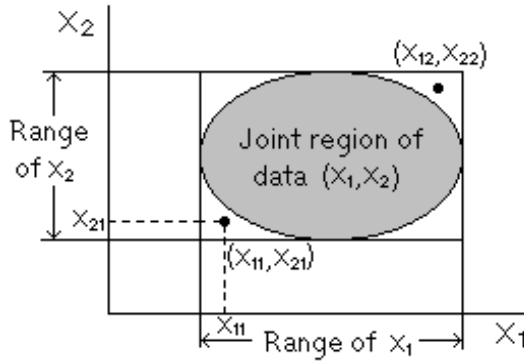


Figure 5.8: An example of extrapolation

determined by (5.296). This problem is sometimes called *hidden extrapolation*, and one should generally check to see whether (5.296) is satisfied before using the model to predict at a new point x_i , $1 \leq i \leq n$.

Some writers have proposed a more stringent definition of the acceptable prediction region [87]. They take this to be the smallest convex set containing $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and call it the *independent variable hull* (IVH). (In the case of simple linear regression this is just the interval $[x_{\min}, x_{\max}]$.) Since this set may be quite difficult to determine explicitly, the ellipsoid (5.296) provides a computable compromise since it contains the IVH and the data points.

5.10 Exercises

- 5.1** Write out the normal equations explicitly for $m = 2$ in (5.1), and find the least squares estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.
- 5.2** Consider a multiple linear regression model with $m \geq 1$ in (5.1). Under the assumption that the errors are i.i.d. $N(0, \sigma^2)$ for a random sample of size n , show that the MLE of σ^2 is given by

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- 5.3** Let $\mathbf{A} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ and $\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Show that $\mathbf{A}^T = \mathbf{A}$, $\mathbf{A}^2 = \mathbf{A}$ and $\mathbf{BA} = \mathbf{0}$.

- 5.4** (Vector differentiation). If $\partial/\partial \boldsymbol{\beta} = ((\partial/\partial \beta_i))_{i=0}^m$, show that

(a) $\partial (\boldsymbol{\beta}^T \mathbf{a}) / \partial \boldsymbol{\beta} = \mathbf{a}$.

(b) $\partial (\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = 2\mathbf{A}\boldsymbol{\beta}$, where \mathbf{A} is symmetric.

5.5 By differentiating $g(\beta) = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle$ with respect to $\beta_j, j = 0, 1, 2, \dots, m$, show that the minimizing value $\hat{\beta}$ satisfies the normal equations $(\mathbf{X}^T \mathbf{X}) \hat{\beta} = \mathbf{X}^T \mathbf{y}$.

5.6 Consider the sample model corresponding to (5.1) as

$$y_i = \beta_0 + \sum_{j=1}^m \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

(a) Find the *least squares function* $g(\beta) = g(\beta_0, \beta_1, \dots, \beta_m) = \sum_{i=1}^n \varepsilon_i^2$.

(b) Differentiate the least squares function found in (a) with respect to β_0 and β_j ($j = 1, 2, \dots, m$) respectively. (Then set them equal to zero.)

(c) Write down the least squares normal equations by simplifying what you found in (b).

5.7 Consider a multiple linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. Prove $\mathbf{X}^T \hat{\varepsilon} = \mathbf{0}$.

5.8 For any linear model, show that

$$\frac{1}{n} \sum_{i=1}^n \text{Var}(\hat{Y}_i) = \frac{1}{n} \text{tr} [\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \sigma^2 = \frac{p\sigma^2}{n}.$$

5.9 Consider the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Show that $SSR = \mathbf{Y}^T \mathbf{H} \mathbf{Y} = \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} = \hat{\mathbf{Y}}^T \mathbf{H}^3 \hat{\mathbf{Y}}$ provided $\bar{\mathbf{Y}} = \mathbf{0}$.

5.10 Consider a multiple linear regression model in (5.1). Let $\mathbf{1}^T = [1, 1, \dots, 1]_{1 \times n}$ and note that $\mathbf{X}^T \mathbf{1}$ is the first column of the design matrix $\mathbf{X}^T \mathbf{X}$. Show that

(a) $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1} = (1, 0, \dots, 0)^T$;

(b) $\mathbf{1}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1} = n$.

[Ref: *The American Statistician*, pp. 47-48 April 1972]

5.11 Suppose that a set of data is given by

y	x_1	x_2
8	4	2
1	9	-8
0	11	-10
5	3	6
3	8	-6
2	5	0
-4	10	-12
11	3	5
-3	7	-2
5	6	-4

(a) Define the design matrix and calculate the inverse of the $\mathbf{X}^T \mathbf{X}$ matrix.

(b) Using the postulated the model for the data, estimate $\beta_i, i = 1, 2$ in the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

(b) Construct the ANOVA table.

(c) Test to determine if the overall regression is statistically significant. Take $\alpha = 5\%$.

(d) Calculate the variances of $\hat{\beta}_1 (= b_1)$, $\hat{\beta}_2 (= b_2)$, and the variance of the predicted value of Y for the point $x_1 = 3, x_2 = 5$.

(e) Calculate R^2 and give a comment.

5.12 Suppose a sample of size $n = 17$ (including repeated runs) was obtained from an experimental study:

Response (y)	x_1	x_2
64, 71	28	71
68, 73	17	43
60, 60, 77, 78	18	27
72, 80	23	32
78, 61, 79, 64	19	90
87	15	40
70	26	7
96	25	96

(a) Fit the model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

(b) Construct the ANOVA and perform the test for lack-of-fit.

(c) Calculate the residuals and plot them. Also give a comment.

(d) Assess the postulated regression model using F, R^2 .

5.13 Let the response Y be a function of three independent variables x_1, x_2 , and x_3 . Suppose that $n = 7$ data points are shown in the table below.

Obs. No.	x_1	x_2	x_3	y
1	-1	-3	1	0
2	-2	0	1	0
3	0	-4	0	1
4	-3	5	-1	1
5	1	-3	-1	2
6	3	5	1	3
7	2	0	-1	3

(a) Fit these data to the model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

(b) Find the predicted value, \hat{y} when $x_1 = 1, x_2 = -3, x_3 = -1$. What is $\hat{y} - y$?

(c) Test $H_0 : \beta_3 = 0$. Take $\alpha = 0.05$.

(d) Construct a 95% confidence interval for the mean response of Y , given $x_1 = 1, x_2 = -3, x_3 = -1$.

(e) Find a 95% prediction interval for Y , given $x_1 = 1, x_2 = -3, x_3 = -1$. Compare this to the answer in (d).

- 5.14** Find the MLEs of the regression coefficients for a multiple regression model $Y = \beta_0 + \sum_{j=1}^m \beta_j x_j + \varepsilon$, $m = 2$. Assume that ε_i 's are independent $N(0, \sigma^2)$.
- 5.15** Show that the fitted plane $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^m \hat{\beta}_j x_j$ passes through the point $(\bar{y}; \bar{x}_1, \bar{x}_2, \dots, \bar{x}_m)$.
- 5.16** Let Y_1, Y_2, \dots, Y_n be observations, which can be postulated by the model

$$Y_i = \beta x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

where x_i 's are fixed constants and ε_i 's are i.i.d. $N(0, \sigma^2)$.

- (a) Find the least square estimator of β .
- (b) Find the MLE of β .
- 5.17** A researcher is interested in studying comparison of the growth rates for bacteria types A and B. The rates of growth were recorded at each of five points of a time period.

Bacteria Type	Time				
	1	2	3	4	5
A	8.0	9.0	9.1	10.2	10.4
B	10.0	10.3	12.2	12.6	13.9

- (a) Using a dummy variable (x_1) for bacteria type, fit the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon.$$

- (b) Plot the data points and graph the growth lines for bacteria types A and B respectively. Note that β_3 is the difference between the slopes of the two lines and represents the interaction between time and bacterial type.
- (c) Do the data indicate sufficient evidence that there exists a difference in the rates of growth for the two types of bacteria?
- (d) Find a 95% confidence interval for the mean growth rate for bacteria type B at time $x_2 = 4$.
- (e) Find a 95% prediction interval for the growth rate Y of bacteria type B at time $x_2 = 4$.
- 5.18** In an experiment, a linear model $Y_i = \beta_0 + \sum_{j=1}^3 \beta_j x_{ij} + \varepsilon_i$, $i = 1, 2, \dots, 8$, is considered. From the data set we have

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 9 \\ 8 \\ 3 \\ 7 \\ 3 \\ 1 \\ 8 \\ 6 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

- (a) Calculate $\mathbf{X}^T \mathbf{X}$, $(\mathbf{X}^T \mathbf{X})^{-1}$, $\mathbf{X}^T \mathbf{Y}$.

- (b) Find the estimated vector $\hat{\beta}$.
- (c) Construct the ANOVA table.
- (d) Test $H_0 : \beta_1 = \beta_3 = 0$. Take $\alpha = 0.10$.
- (e) Compute the R^2 for the reduced model $Y = \beta_0 + \beta_3 x_1 x_2 + \varepsilon$.

5.19 The cloud point of a liquid is a measure of the degree of crystallization in a stock that can be measured by the refractive index. It has been suggested that the percentage of I-8 in the base stock is an excellent predictor of cloud point. The following data were collected on stocks with a known percentage of I-8 [27].

% I-8	Cloud Point, Y	% I-8	Cloud Point, Y
0	22.1	2	26.1
1	24.5	4	28.5
2	26.0	6	30.3
3	26.8	8	31.5
4	28.2	10	33.1
5	28.9	0	22.8
6	30.0	3	27.3
7	30.4	6	29.8
8	31.4	9	31.8
0	21.9		

- (a) Fit the data using the second-order model: $Y = \beta_0 + \beta_1 x + \beta_{21} x^2 + \varepsilon$.
- (b) Is the overall regression model in (a) significant? Take $\alpha = 0.05$.
- (c) Construct the ANOVA table and perform the test of lack of fit for the model in (a).
- (d) Fit the data using the first-order model: $Y = \beta_0 + \beta_1 x + \varepsilon$.
- (e) Use the residuals from the result in (d), make a conclusion whether the first-order model would have been sufficient.

5.20 Suppose we have a parameter vector $\beta = [\beta_1, \beta_2, \beta_3, \beta_4]^T$ to be estimated in a linear model. Suppose one wishes to test the following hypotheses. Specify the matrices \mathbf{C} and \mathbf{b} in the linear hypothesis of the form $H_0 : \mathbf{C}\beta = \mathbf{b}$.

- (a) $H_0 : \beta_1 = \beta_3, \beta_2 = \beta_4$.
- (b) $H_0 : \beta_1 = \beta_3 = 0, \beta_2 = \beta_4 = 0$.
- (c) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

5.21 Let $Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ ($i = 1, 2, \dots, I, j = 1, 2, \dots, J$), where $\sum_i d_i \beta_i = 0$ ($\sum_i d_i \neq 0$) and $E(\varepsilon_{ij}) = 0$ for all i, j . Using the method of Lagrange multipliers find the least squares estimates of μ and β_i [104].

[Hint: Show that the Lagrange multiplier is zero.]

Chapter 6

Residuals, Diagnostics and Transformations

6.1 Introduction

As for simple linear regression, it is important to check the validity of a proposed model. Of course, the statistics such as R^2 , t and F tests are important methods for doing this. However, these tests were all derived under the assumption that (5.1) was the true model and the assumptions about normality and constant variance of errors were valid. In general, as we noted in Chapter 3, these assumptions can usually only be tested after the model has been fitted, and as was done there, the examination of residuals is an important tool for doing this.

In addition, it is important to examine the data itself for effects on the fit, because even if the model is correct, the available sample of observations may make it difficult to obtain good estimates of the parameters. As we have shown in Examples 5.24 and 5.27 strong multicollinearity can confound the parameter estimates, even if the overall model fits the data well. In addition, it is important to examine the effect of the observations on the fit, since particular cases may unduly affect parameter estimates. Traditionally, this was done by looking for outliers in the fitted values $\hat{\mathbf{y}}$. Clearly, a large residual at a given observation y_i suggests that there may be a problem with the model. However, the opposite can happen as well, a discrepant point may be masked by the fact that it is highly influential on the fit. Over the past 20 years this subject has gained increasing attention among statisticians and such examination should become a routine part of modern regression analysis.

In this Chapter, we will develop a number of tools, generalizing those we discussed in Chapter 3 and a number of new regression diagnostics to look for influential data points. Finally, if these procedures discover apparent model violations, we discuss a variety of techniques for variable transformations and methods for correction of non-constant variance.

6.2 Residuals

In this section we establish a number of properties of the residuals from the least squares fit of (5.1). We begin by summarizing some of those established in Chapter 3 and Chapter 5 and discuss several new ones which are important in justifying the various residual plots discussed in Section 6.3.

6.2.1 Properties of $\hat{\varepsilon}$

Recall from Chapter 5 that the fitted values $\hat{\mathbf{y}}$ are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y} \quad (6.1)$$

where, as before, \mathbf{H} is called the “hat matrix” - so-called because it transforms \mathbf{y} into $\hat{\mathbf{y}}$ (y-hat). Using (6.1) the residuals $\hat{\varepsilon}$ are defined by

$$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \quad (6.2)$$

If (5.1) is true, then $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, so that (using $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$)

$$\begin{aligned} \hat{\varepsilon} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}. \end{aligned} \quad (6.3)$$

From (6.3) it follows that

$$E(\hat{\varepsilon}) = (\mathbf{I} - \mathbf{H})E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad (6.4)$$

and using the fact that $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent

$$\begin{aligned} \boldsymbol{\Sigma}(\hat{\varepsilon}) &= (\mathbf{I} - \mathbf{H})^T \boldsymbol{\Sigma}(\boldsymbol{\varepsilon}) (\mathbf{I} - \mathbf{H}) \\ &= \sigma^2 (\mathbf{I} - \mathbf{H})^2 = \sigma^2 (\mathbf{I} - \mathbf{H}). \end{aligned} \quad (6.5)$$

Moreover, if $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, then it follows from (6.3) that $\hat{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 (\mathbf{I} - \mathbf{H}))$. From (6.5) it follows that

$$\text{Var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii}) \quad (6.6)$$

so that (h_{ii} is the i -th diagonal element of \mathbf{H})

$$\hat{\varepsilon}_i \sim N(0, \sigma^2 (1 - h_{ii})) \quad (6.7)$$

and

$$\text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 h_{ij}. \quad (6.8)$$

As noted in Chapter 3, it is generally preferable to standardize the residuals in some ways. Since $\text{Var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii})$, $r_i = \hat{\varepsilon}_i / \sigma \sqrt{1 - h_{ii}}$ has $\text{Var}(r_i) = 1$ and if we estimate σ by $s = \sqrt{SSE / (n - m - 1)}$, then

$$\hat{r}_i = \frac{\hat{\varepsilon}_i}{s \sqrt{1 - h_{ii}}}, \quad 1 \leq i \leq n, \quad (6.9)$$

is called the *internally studentized residual*. On the other hand, if σ^2 is estimated from the least squares fit by omitting the i -th observation, then, denoting this estimate of σ by $\hat{\sigma}_{(-i)}$

$$\hat{t}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}, \quad 1 \leq i \leq n \quad (6.10)$$

is called the *externally studentized residual*. In computer programs \hat{t}_i is frequently referred to as RSTUDENT. We leave it as an exercise for the reader to show that these formulas reduce to those given in Chapter 3 for simple linear regression.

From (6.9)-(6.10) we see that if h_{ii} is small, then for large n , we would expect that $\hat{\varepsilon}_i$, \hat{r}_i and \hat{t}_i should all be approximately $N(0, 1)$ and so should behave roughly the same. For small n (say $n < 20$) and/or h_{ii} close to one, then either \hat{r}_i or \hat{t}_i is to be preferred over $\hat{\varepsilon}_i$ and current statistical practice seems to favor using \hat{t}_i , particularly if the *leverage* h_{ii} of y_i is large.

In fact, as we will show, h_{ii} plays a major role in many of the currently used regression diagnostics. For future reference we now discuss some of its basic properties.

6.2.2 The Leverage h_{ii}

Since $\text{Var}(\hat{\beta}_i) \geq 0$, it then follows from (6.6) that $h_{ii} \leq 1$. In addition, since $\mathbf{H}^2 = \mathbf{H}$, then

$$h_{ii} = \sum_{k=1}^n h_{ik}h_{ki} = \sum_{k=1}^n (h_{ik})^2. \quad (6.11)$$

Thus, $h_{ii} > 0$. (One can actually show the stronger inequality $h_{ii} \geq 1/n$ [20].)

The importance of h_{ii} comes from the following observation. Using the fact that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$

$$\hat{y}_i = \sum_{j=1}^n h_{ij}y_j = h_{ii}y_i + \sum_{i \neq j} h_{ij}y_j. \quad (6.12)$$

Now if $h_{ii}, i \neq j$ are “small”, then it follows from (6.12) that $\hat{y}_i \simeq h_{ii}y_i$ so that h_{ii} measures the effect that y_i has on determining the fitted value \hat{y}_i . If $h_{ii} \simeq 1$, then $\hat{y}_i \simeq y_i$ and the model is forced to fit the model through the point (\mathbf{x}_i, y_i) even if the model is not valid there. These points with “large” h_{ii} are called *high leverage points* and should be singled out for further investigation. Note from (6.12) that if y_i is a *high leverage* point, then $\hat{\varepsilon}_i$ could be large, but \hat{t}_i “small”, with the leverage of y_i masking the effect of a poor model fit at y_i . This masking effect is important to be aware of when using plots and/or other diagnostics to evaluate the model

A further issue concerns what values of h_{ii} should be considered large. A current popular heuristic for doing this results from the following argument. We first observe that

$$\sum_{i=1}^n h_{ii} = \text{tr}(\mathbf{H}) = \text{tr}(\mathbf{I}_{m+1}) = m+1 \quad (6.13)$$

as shown in the proof of Theorem 5.4. Since $h_{ii} > 0$, then $(m+1)/n$ is the average size of h_{ii} and Belsley Kuh and Welsch (BKW) [8] recommend defining h_{ii} as large, if $h_{ii} > 2(m+1)/n$. In general, modern regression software will flag points with $|\text{RSTUDENT}| > 2$ and or $h_{ii} > 2(m+1)/n$ as points to be considered for further consideration. We will return to this matter shortly.

6.3 Residual Plots

6.3.1 Normal Plots

In general, as in Chapter 3 various residual plots can be used to check for the validity of model assumptions. To check for normality one can use histograms or normal plots as for simple linear regression (SLR). Since these plots do not depend on the number of independent variables they are obtained as for SLR.

Although there are several analytic tests for normality [104, 27], they are quite complicated and seem not to be widely used.

6.3.2 Variable Plots

To check for model violations in the functional form of $E(Y_{\mathbf{x}})$ and/or non-constant variance one typically uses the following plots:

- (i) Plots of $\hat{\varepsilon}_i$, \hat{r}_i or \hat{t}_i against \mathbf{x}_j , $1 \leq j \leq m$, the j -th column of \mathbf{X} .
- (ii) Plots of $\hat{\varepsilon}_i$, \hat{r}_i or \hat{t}_i versus $\hat{\mathbf{y}}$.
- (iii) Partial plots.
- (iv) Other plots as indicated.

To justify (i) and (ii) we note the following. From the least squares equations we have $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$. Since $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$, $\mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}^T\hat{\boldsymbol{\varepsilon}} = 0$. From (3.162) if we regress $\hat{\boldsymbol{\varepsilon}}$ on \mathbf{x}_j the slope of the regression line is

$$\hat{\beta}_j^* = \langle \mathbf{x}_j, \hat{\boldsymbol{\varepsilon}} \rangle / \|\mathbf{x}_j\|^2. \quad (6.14)$$

Since the j -th row of \mathbf{X}^T is the j -th column of \mathbf{X} , it follows from (6.14) that if $\langle \mathbf{x}_j, \hat{\boldsymbol{\varepsilon}} \rangle = 0$ that $\hat{\beta}_j^* = 0$, $1 \leq j \leq m$. From this argument we see that if (5.1) is the true model then plotting $\hat{\varepsilon}_j$, \hat{r}_j or \hat{t}_j against x_j should give a random scatter of points about the “ x ”-axis.

In using $\hat{\boldsymbol{\varepsilon}}$, $\hat{\mathbf{r}}$ or $\hat{\mathbf{t}}$, these points should lie roughly in a band between ± 2 on the $\hat{\boldsymbol{\varepsilon}}$ axis. Deviations from this suggest nonlinearity of the model in \mathbf{x} ; and/or nonconstant variance.

For (ii) we showed in Theorem 5.6 that $\sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i = 0$, so that the slope of the least squares line (through the origin) of $\hat{\boldsymbol{\varepsilon}}$ against $\hat{\mathbf{y}}$ is

$$\hat{\alpha} = \langle \hat{\boldsymbol{\varepsilon}}, \hat{\mathbf{y}} \rangle / \|\hat{\mathbf{y}}\|^2 = 0. \quad (6.15)$$

So again, if the model is true, plots of $\hat{\boldsymbol{\varepsilon}}$ against $\hat{\mathbf{y}}$ should be randomly scattered about the abscissa as in indicated in Figure 3.17. Deviations from this pattern again suggest nonconstant variance if there is a funnel shape, or nonlinearity if there are “trends” in the plot. Some authors have suggested that variance inequality is more easily seen by plotting $|\hat{\varepsilon}_i|$ against \hat{y}_i , $1 \leq i \leq n$.

6.3.3 Partial Plots

Although plots of $\hat{\varepsilon}_i$ versus \mathbf{x}_j and $\hat{\varepsilon}$ against $\hat{\mathbf{y}}$ can indicate defects in the model, they may not be informative as to what these defects are. In SLR plotting $\hat{\varepsilon}_i, 1 \leq i \leq n$, against $x_i, 1 \leq i \leq n$, can be used to detect deviations in functional form from linearity. However, in multiple linear regression (MLR) such plots, as with scatter plots, can be confused by the fact that $\hat{\varepsilon}$ depends on all of the predictors, so does not necessarily isolate the effects of a given variable with the effects other variables removed. To remedy this, a number plots have been proposed which better isolate the behavior of the j -th variable. Those plots may be thought of as substitutes for scatter plots in SLR. We discuss two such plots [87]:

- (i) Partial residual plots;
- (ii) Partial regression plots (also called *added variable plots*).

Partial Residual Plots

For partial residual plots we define

$$\hat{\varepsilon}_j^* = \hat{\varepsilon} + \hat{\beta}_j \mathbf{x}_j \quad (6.16)$$

where $\hat{\varepsilon}$ is the residual vector from the least squares fit of (5.1) and $\hat{\beta}_j$ is the least squares estimate of β_j . For the partial residual plot we plot $\hat{\varepsilon}_j^*$ against $\mathbf{x}_j, 1 \leq j \leq m$. If the model is correct, then this plot should display a random scatter about the line with slope $\hat{\beta}_j$. Deviations can be used to detect violations in the assumption of linearity in \mathbf{x}_j .

To see this, notice first that the relation between $\hat{\varepsilon}_j^*$ and \mathbf{x}_j is of the form of the SLR model for regression through the origin. If the model is true, then $\hat{\varepsilon}_i, 1 \leq i \leq n$, should scatter about the x -axis and so we can consider regressing $\hat{\varepsilon}_j^*$ against \mathbf{x}_j . From (3.18) the slope of the least squares line is given by

$$\hat{\gamma}_j = \frac{\langle \hat{\varepsilon}_j^*, \mathbf{x}_j \rangle}{\langle \mathbf{x}_j, \mathbf{x}_j \rangle} = \frac{\langle \hat{\varepsilon} + \hat{\beta}_j \mathbf{x}_j, \mathbf{x}_j \rangle}{\langle \mathbf{x}_j, \mathbf{x}_j \rangle}. \quad (6.17)$$

Now, $\langle \hat{\varepsilon} + \hat{\beta}_j \mathbf{x}_j, \mathbf{x}_j \rangle = \langle \hat{\varepsilon}, \mathbf{x}_j \rangle + \hat{\beta}_j \langle \mathbf{x}_j, \mathbf{x}_j \rangle$. As before, $\langle \hat{\varepsilon}, \mathbf{x}_j \rangle = 0$, so that the numerator in (6.17) is $\hat{\beta}_j \langle \mathbf{x}_j, \mathbf{x}_j \rangle$. Thus,

$$\hat{\gamma}_j = \hat{\beta}_j. \quad (6.18)$$

To illustrate some of the ideas concerning the various residual plots we begin with an examination of the housing price model discussed in Example 5.12. Recall there that the square footage and age were significant in explaining house prices with an $R^2 = 0.959$. Further, we found that $t_1 = 16.63$ and $t_2 = -4.74$ both of which are significant at $< 0.1\%$ level. However, we need to consider whether the assumptions under which these results were derived are valid. To do this we consider a number of residual plots as suggested above. As in Chapter 3, we display normal residual plots, plots of various histograms and various types of residual plots against fitted values and variables. We begin by considering the difference in using $\hat{\varepsilon}_i, \hat{r}_i$ and \hat{t}_i .

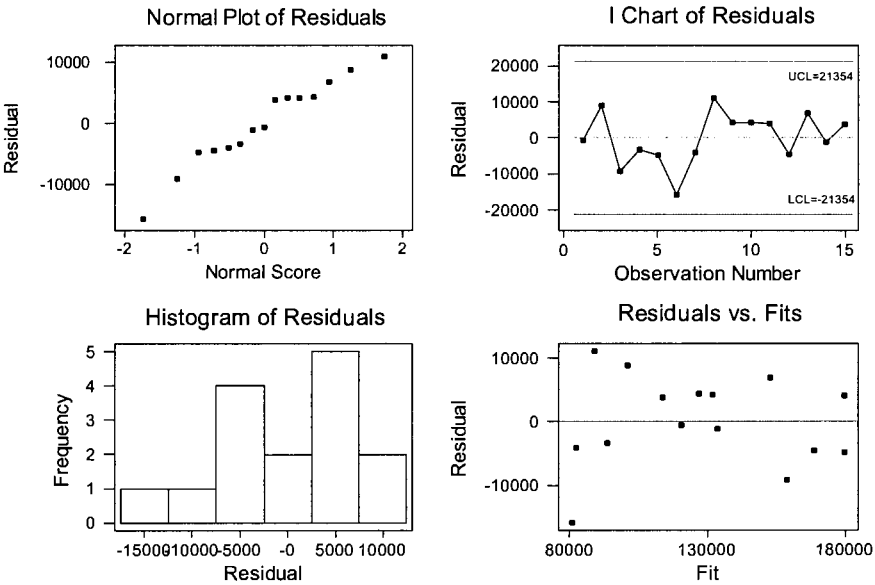


Figure 6.1: Housing price data and ordinary residuals $\hat{\epsilon}_i$

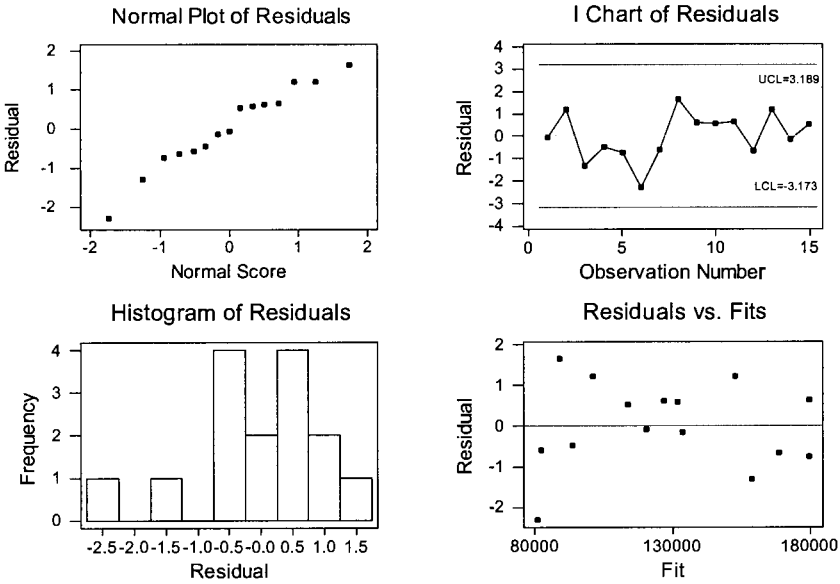


Figure 6.2: Plots of standardized residuals \hat{r}_i in Housing price data

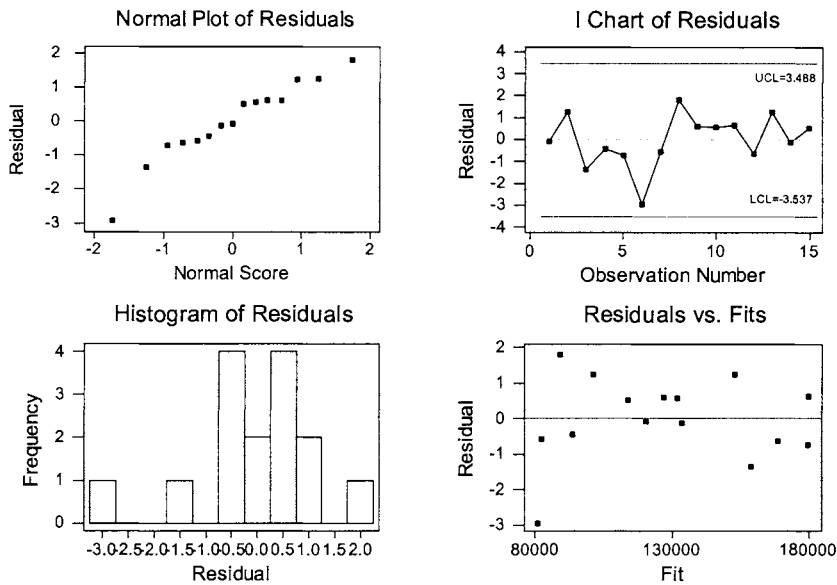


Figure 6.3: Plots of studentized residuals \hat{t}_i in Housing price data

In Figure 6.1 we show these plots using the ordinary residuals $\hat{\epsilon}_i$, in Figure 6.2 there are plots using the standardized residuals \hat{r}_i and in Figure 6.3 the plots using the studentized residuals \hat{t}_i . A table of the various residuals and fitted values and their leverages are given in Table 6.1.

Table 6.1 Residuals and Leverage for Housing Data

Obs. No.	Fitted (\hat{y}_i)	Resi ($\hat{\epsilon}_i$)	S Resi (\hat{r}_i)	T Resi (\hat{t}_i)	Hi (h_{ii})
1	120,572	-572.1	-0.07917	-0.0758	0.13335
2	101,123	8876.8	1.21117	1.2377	0.10868
3	159,137	-9137.4	-1.31776	-1.3642	0.20219
4	93,338	-3337.5	-0.46350	-0.4478	0.13963
5	179,859	-4858.5	-0.74759	-0.7330	0.29917
6	80,796	-15795.7	-2.29877	-2.9420	0.21654
7	82,068	-4067.9	-0.58926	-0.5725	0.20924
8	88,975	11025.1	1.64673	1.7921	0.25621
9	127,116	4383.9	0.61565	0.5990	0.15862
10	131,872	4227.7	0.58567	0.5689	0.13536
11	179,859	4141.5	0.63726	0.6207	0.29917
12	169,043	-4543.4	-0.66239	-0.6461	0.21935
13	153,109	6891.1	1.20659	1.2324	0.45877
14	133,569	-1068.6	-0.14352	-0.1375	0.08021
15	113,665	3834.9	0.51602	0.4996	0.08353

Notice that the normal plots, I Charts and residual plots against the fitted values \hat{y}_i are all similar in appearance. Overall, there is nothing remarkably out of line from our basic normality assumptions except the residual corresponding to observation 6 which clearly appears as an outlier with a substantial underpricing for its size and age. However, its leverage $h_6 = 0.212$ which is less than $(2 \times 3) / 15 = 0.4$ the cut off value suggested by BKW so does not appear to be overly influencing the overall fit. Perhaps there is something special about that house such as its condition relative to the others. If such information is available it should be considered before omitting this observation from the model.

Although the residual plots look similar, there is some difference in the histograms. The histogram for $\hat{\varepsilon}_i$ is clearly non-normal in appearance, possibly reflecting the unequal variances and dependence of $\hat{\varepsilon}_i$ on \mathbf{X} , while those for \hat{r}_i and \hat{t}_i are more symmetric in appearance with the plot for \hat{t}_i being most normal in appearance. Overall it appears that these plots reveal no substantial deviations from the basic normality assumptions.

Further confirmation of these observations is given in Figures 6.4-6.5 which show t_i plotted against square footage and age respectively. Again, the only feature which stands out is observation 6.

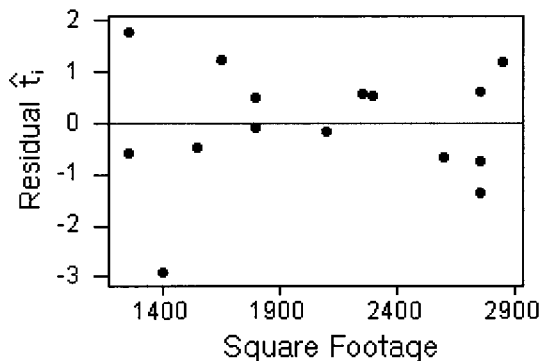


Figure 6.4: Plot of \hat{t}_i versus square footage (ft)

Finally, to see if the outlying observation might be accounted for by some nonlinear behavior in the predictors we present partial residual plots for the partial residuals ε_1^* , ε_2^* against x_1 and x_2 respectively in Figures 6.6-6.7. They are striking. Both plots are almost perfect straight lines and should be contrasted to the scatter plots given in Figures 5.2-5.3. In particular, the partial plot for age shows a clear decrease, for increasing age, in contrast to the scatter plot Figure 5.3. To verify (6.16) we regressed ε_1^* on \mathbf{x}_1 and ε_2^* on \mathbf{x}_2 and the estimated slopes were

$$\hat{\beta}_1^* = 60.5890 \quad \text{and} \quad \hat{\beta}_2^* = -1726.0$$

which are essentially identical to the least squares estimates

$$\hat{\beta}_1 = 60.589 \quad \text{and} \quad \hat{\beta}_2 = -1726.8.$$

In order to keep the number of plots within reason, we generally just display plots of \hat{t}_i against \hat{y}_i and the corresponding normal plots and histograms for model checking

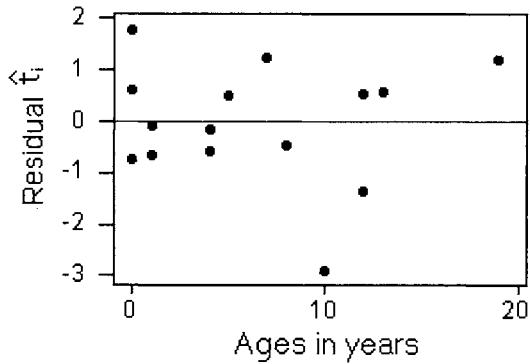


Figure 6.5: Plot of \hat{t}_i versus age in years

purposes and residual plots against variables when warranted. The reader is encouraged to obtain other plots as necessary.

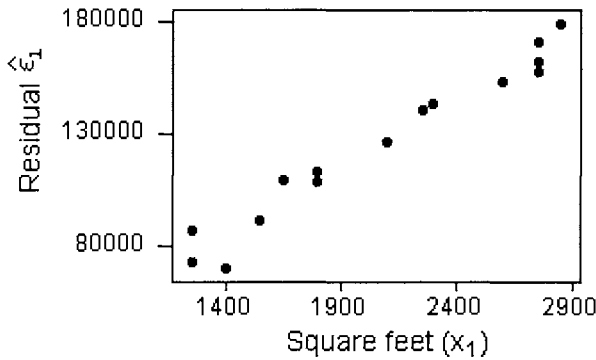


Figure 6.6: Partial residual plot of ϵ_1^* against x_1

Example 6.1 (Drink delivery data) In Example 5.14 we indicated that the time to delivery was quite well explained by the linear model (5.99). To further check the validity of the proposed model we present a table of residuals and leverage values in Table 6.2 and residual plots of \hat{t}_i against \hat{y}_i in Figure 6.8. Again, these plots are reasonably in accord with our normality assumptions except for observation 9 and observation 22. As we can see from Table 6.2 $h_9 = 0.4983$ and $h_{22} = 0.3916$ so both exceed the cutoff value $(2 \times 3) / 25 = 0.24$ hence appear to be high leverage points which may be affecting the overall fit of the model. In addition, the histogram of \hat{t}_i looks skewed and somewhat non-normal. Similar comments hold for the normal plot. Consequently, we need to be concerned that the linear model (5.99) is not telling the whole story. One needs to account for observations 9 and 22 and the possibility that another error model is more appropriate than normal. In fact in the OzDASL¹ it was suggested that a *gamma distribution* was

¹Australian Data and Story Library, WEB site is <http://www.statsci.org>

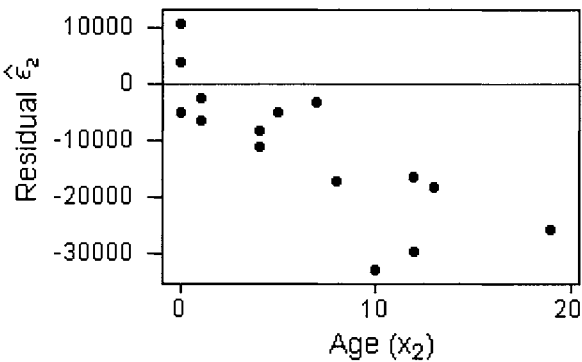


Figure 6.7: Partial residual plot of ε_2^* against x_2

a more appropriate error distribution than the normal. Such distributions are members of the exponential family and can be fit using a modification of least squares [87]. (We shall briefly discuss this topic in Section 7.7 but the details are beyond the scope of this text.)

Table 6.2 Residuals and Leverage for Drink Delivery Data

Obs. No.	Fitted (\hat{y}_i)	Resi ($\hat{\varepsilon}_i$)	S Resi (\hat{r}_i)	T Resi (\hat{t}_i)	Hi (h_{ii})
1	21.708	-5.02808	-1.62768	-1.69563	0.10180
2	10.354	1.14639	0.36484	0.35754	0.07070
3	12.080	-0.04979	-0.01609	-0.01572	0.09874
4	9.956	4.92435	1.57972	1.63916	0.08538
5	14.194	-0.44440	-0.14176	-0.13856	0.07501
6	18.400	-0.28957	-0.09081	-0.08874	0.04287
7	7.155	0.84462	0.27042	0.26465	0.08180
8	16.673	1.15660	0.36672	0.35939	0.06373
9	71.82	7.41971	3.21376	4.31078	0.49829
10	19.124	2.37641	0.81325	0.80678	0.19630
11	38.093	2.23749	0.71808	0.70994	0.08613
12	21.593	-0.59304	-0.19326	-0.18897	0.11366
13	12.473	1.02701	0.32518	0.31847	0.06113
14	18.682	1.06754	0.34114	0.33418	0.07824
15	23.329	0.67120	0.21029	0.20566	0.04111
16	29.663	-0.66293	-0.22270	-0.21783	0.16594
17	14.914	0.43636	0.13804	0.13492	0.05943
18	15.551	3.44862	1.11295	1.11933	0.09626
19	7.707	1.79319	0.57877	0.56981	0.09645
20	40.888	-5.78797	-1.87355	-1.99668	0.10169
21	20.514	-2.61418	-0.87784	-0.87309	0.16528
22	56.007	-3.68653	-1.45000	-1.48962	0.39158
23	23.358	-4.60757	-1.44369	-1.48247	0.04126
24	24.403	-4.57285	-1.49606	-1.54222	0.12061
25	10.963	-0.21258	-0.06751	-0.06596	0.06664

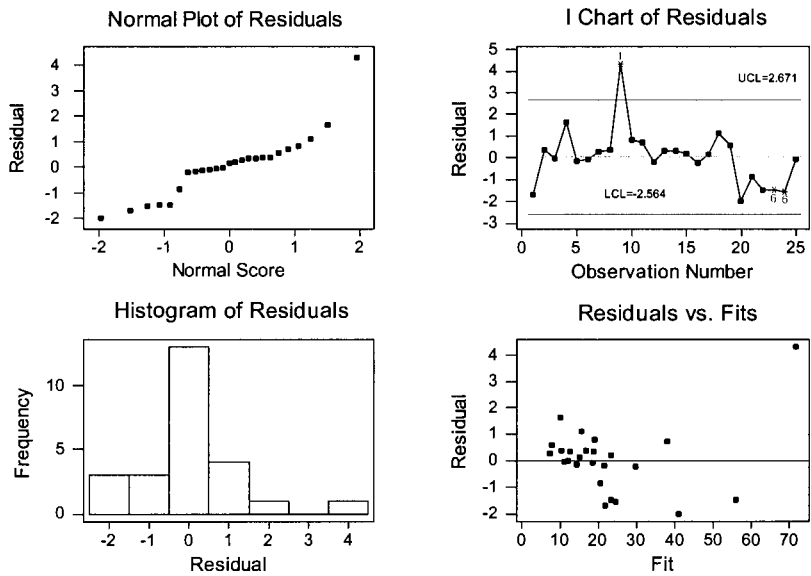


Figure 6.8: Plots of RSTUDENT (\hat{t}_i) versus \hat{y}_i for delivery data

Example 6.2 (Birth weight data) In Example 5.15 we concluded that the birth weight data could reasonably be described by the model (5.102) which represents two parallel lines. However, because there is a substantial amount of unexplained variation in the data ($R^2 = 0.66$) one might be concerned that other variables might be useful for explaining the data. To examine this possibility we first examine the residual plots for \hat{t}_i . Figure 6.9 shows a rather random distribution of residuals and all lie between the ± 2 cut-off values. The histogram has a very normal appearance, although some skewness is indicated in the normal plot. Nothing in these plots suggests violation of the normality assumptions. For further verification we constructed the partial residual plots and they are shown in Figures 6.10-6.11. As in Example 6.1 these plots isolate the effect much more dramatically than the scatter plots. In Figure 6.11 the data generally fall on a line but there is considerable variability at some points. Since x_2 is a dummy 0-1 variable, the partial residual plot consists of two vertical lines, with male values clearly higher on average than the females. Notice however, that the variability of the male residuals appears larger than that for females.

Finally, we fitted the data in Figures 6.10-6.11 by least squares. For $\hat{\varepsilon}_1^*$ the slope was 119 while for $\hat{\varepsilon}_2^*$ the slope is 182.2. Both of these are virtually identical to the least squares estimates $\hat{\beta}_1 = 118.67$ and $\hat{\beta}_2 = 166.28$. These observations indicate that the normality assumptions appear to be valid and the model (5.103) an adequate representation of the data.

Table 6.3 Residuals and Leverage for Birth Weight Data

No.	Fitted (\hat{y}_i)	Resi ($\hat{\varepsilon}_i$)	S Resi (\hat{r}_i)	T Resi (\hat{t}_i)	Hi (h_{ii})
1	3225.95	-257.949	-1.5161	-1.5679	0.1204
2	2988.61	-193.610	-1.1156	-1.1225	0.0848
3	3225.95	-62.949	-0.3700	-0.3623	0.1204
4	2632.60	342.399	2.1533	2.3805	0.2316
5	2751.27	-126.271	-0.7577	-0.7497	0.1560
6	2869.94	-22.940	-0.1338	-0.1307	0.1071
7	3344.62	-52.619	-0.3200	-0.3130	0.1783
8	3225.95	247.051	1.4521	1.4941	0.1204
9	2869.94	-241.940	-1.4114	-1.4477	0.1071
10	2988.61	187.390	1.0798	1.0843	0.0848
11	3225.95	195.051	1.1464	1.1556	0.1204
12	2988.61	-13.610	-0.0784	-0.0766	0.0848
13	3059.67	257.330	1.4987	1.5477	0.1042
14	2584.99	144.008	0.8789	0.8740	0.1843
15	3059.67	-124.670	-0.7261	-0.7177	0.1042
16	2822.33	-68.331	-0.3950	-0.3870	0.0908
17	3297.01	-87.010	-0.5446	-0.5353	0.2243
18	2941.00	-124.001	-0.7143	-0.7057	0.0842
19	3059.67	66.330	0.3863	0.3784	0.1042
20	2703.66	-164.661	-0.9699	-0.9685	0.1242
21	2584.99	-172.992	-1.0558	-1.0589	0.1843
22	2822.33	168.669	0.9751	0.9739	0.0908
23	2941.00	-66.001	-0.3802	-0.3723	0.0842
24	3059.67	171.330	0.9979	0.9978	0.1042

Example 6.3 (Longley data) As we have seen in Example 5.17 total employment 1947-1962 could be explained by the six predictors x_1 - x_6 discussed there, but the apparent multicollinearity in the data made it difficult to identify the important variables. Using a stepwise regression analysis it was suggested that an adequate model could be obtained using x_2 and x_3 . To further consider the validity of these models we examine a variety of residual plots.

In Table 6.4 we display the residuals and leverages from the full model. There appear to be no outliers, although observation 10 (1956) may be somewhat suspicious, and no apparent high leverage points. The corresponding residuals \hat{t}_i are given in Figure 6.12. The normal plot and histogram appear normal but there appears to be some autocorrelation in the residuals.

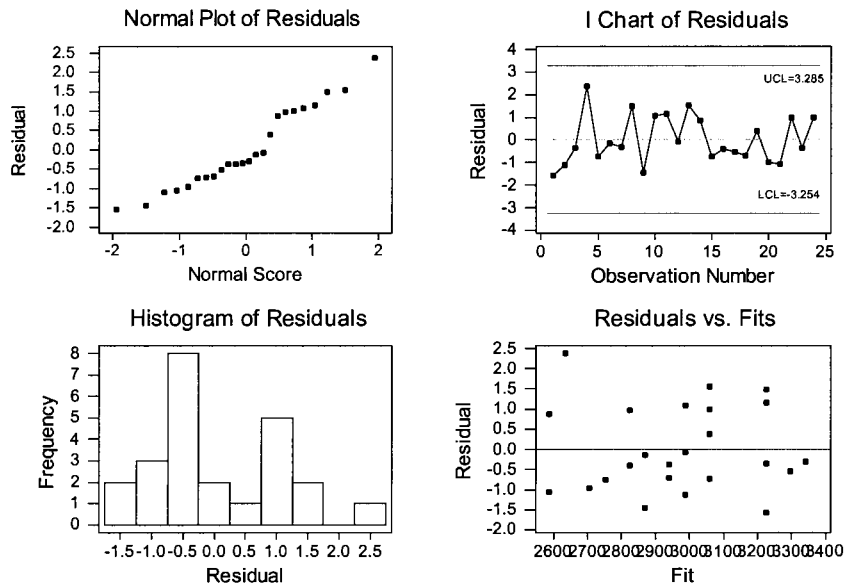


Figure 6.9: Plots of RSTUDENT (\hat{t}_i) versus \hat{y}_i for birth weight data

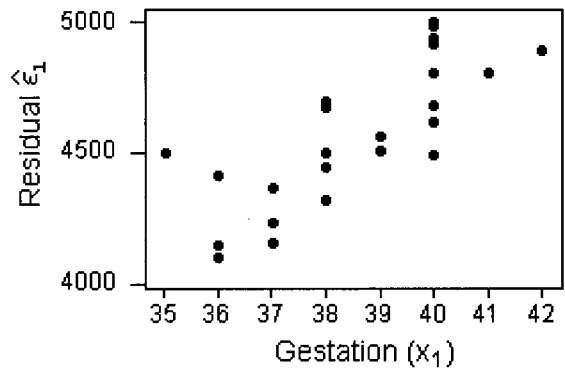


Figure 6.10: Partial residual plot for x_1 (gestation)

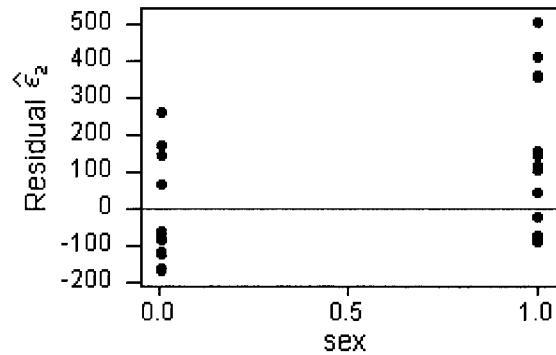


Figure 6.11: Partial residual plot for x_2 (sex)

Table 6.4 Residuals and Leverage for Longley data

No.	Fitted (\hat{y}_i)	Resi ($\hat{\varepsilon}_i$)	S Resi (\hat{r}_i)	T Resi (\hat{t}_i)	Hi (h_{ii})
1	60.0557	0.2673	1.1639	1.1906	0.4336
2	61.2200	-0.0980	-0.4834	-0.4618	0.5587
3	60.1270	0.0440	0.1805	0.1705	0.3610
4	61.5934	-0.4064	-1.6909	-1.9300	0.3798
5	62.9060	0.3150	1.6639	1.8853	0.6152
6	63.8874	-0.2484	-1.0252	-0.0285	0.3694
7	65.1597	-0.1707	-0.7809	-0.7625	0.4871
8	63.7767	-0.0157	-0.0731	-0.0689	0.5046
9	65.9988	0.0202	0.0902	0.0851	0.4596
10	67.4005	0.4565	1.8287	2.1750	0.3308
11	68.1887	-0.0197	-0.0808	-0.0762	0.3607
12	66.5520	-0.0390	-0.1778	-0.1679	0.4840
13	68.8042	-0.1492	-0.6154	-0.5928	0.3688
14	69.6493	-0.0853	-0.3181	-0.3016	0.2283
15	68.9915	0.3395	1.4050	1.4992	0.3730
16	70.7612	-0.2102	-1.2282	-1.2692	0.6854

In Figure 6.13 are shown the residual plots for \hat{t}_i for the “best” model selected by the stepwise procedure. Again, the normal plot and histogram look reasonably normal but now observation 10 appears as an outlier and the I Chart and residual plot display much more pronounced autocorrelation. The overall impression is that the effect of time has not been properly accounted in the reduced model. We will return to this matter later.

Partial Regression (Leverage) Plots

Partial residual plots have been criticized by some statisticians for overestimating the effect of \mathbf{x}_j on the fit and other plots have been suggested as alternatives. An important plot, called a *partial regression* or *added variable plot* is widely advocated as an alternative (or complement) to partial residual plots. To motivate these plots, consider the problem

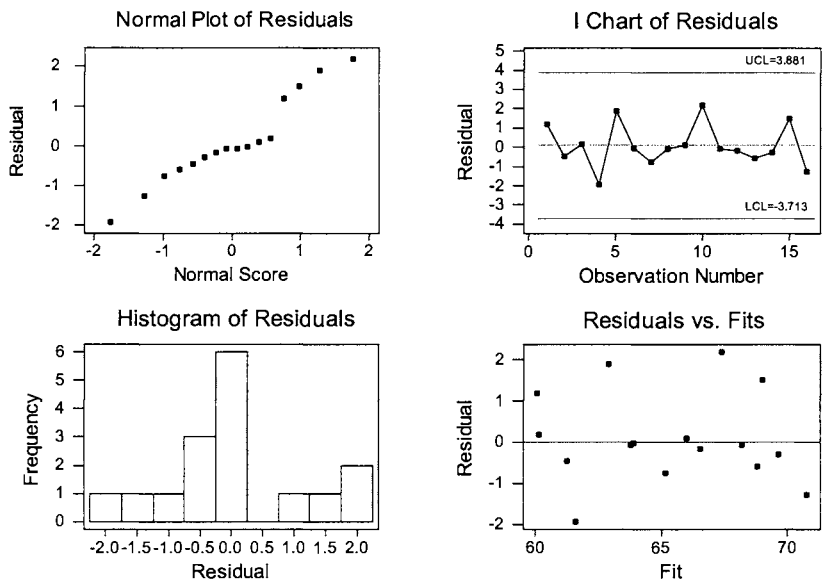


Figure 6.12: Plots of residuals for full Longley data

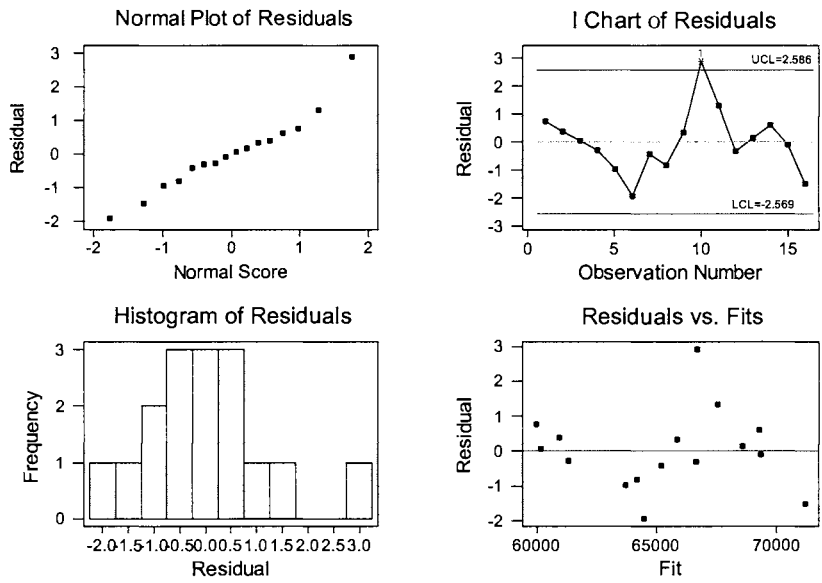


Figure 6.13: Plots of residuals \hat{t}_i for x_2, x_3 (best model)

of deciding whether to add a new explanatory variable to the model (5.1) and we wish to estimate its effect. This will be determined by fitting the augmented model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{w} + \boldsymbol{\varepsilon} \quad (6.19)$$

obtained by appending \mathbf{w} to the design matrix. Hence we can write (6.19) in partitioned form as

$$\begin{aligned} \mathbf{Y} &= [\mathbf{X} \mid \mathbf{w}] \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \end{pmatrix} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}_w \boldsymbol{\beta}_w = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{w} + \boldsymbol{\varepsilon} \end{aligned} \quad (6.20)$$

The model (6.19) is fit by least squares giving the estimate $(\hat{\boldsymbol{\alpha}} \mid \hat{\gamma})^T$ where $\hat{\boldsymbol{\alpha}}$ is the revised estimate for $\boldsymbol{\beta}$ in (6.20). (Generally, $\hat{\boldsymbol{\alpha}} \neq \hat{\boldsymbol{\beta}}$, unless \mathbf{w} is orthogonal to the columns of \mathbf{X}). We now derive a formula for $\hat{\gamma}$. From (6.20)

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}_w^T \mathbf{X}_w)^{-1} \mathbf{X}_w^T \hat{\mathbf{y}} \quad (6.21)$$

or equivalently $\hat{\boldsymbol{\beta}}_w$ satisfies

$$(\mathbf{X}_w^T \mathbf{X}_w) \hat{\boldsymbol{\beta}}_w = \mathbf{X}_w^T \hat{\mathbf{y}}. \quad (6.22)$$

Now from (6.22)

$$\mathbf{X}_w^T \mathbf{X}_w = \begin{bmatrix} \mathbf{X}^T \\ \mathbf{w}^T \end{bmatrix} [\mathbf{X} \mid \mathbf{w}] = \begin{bmatrix} \mathbf{X}^T \mathbf{X} + \mathbf{X}^T \mathbf{w} \\ \mathbf{w}^T \mathbf{X} + \mathbf{w}^T \mathbf{w} \end{bmatrix}. \quad (6.23)$$

Thus, $\hat{\boldsymbol{\alpha}}$ and $\hat{\gamma}$ satisfy

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\alpha}} + (\mathbf{X}^T \mathbf{w}) \hat{\gamma} = \mathbf{X}^T \mathbf{y}, \quad (6.24)$$

$$\mathbf{w}^T \mathbf{X} \hat{\boldsymbol{\alpha}} + (\mathbf{w}^T \mathbf{w}) \hat{\gamma} = \mathbf{w}^T \mathbf{y}. \quad (6.25)$$

From (6.24)

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{w} \hat{\gamma} \quad (6.26)$$

and substituting (6.25) into (6.24) gives

$$\mathbf{w}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} - \mathbf{w}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{w} \hat{\gamma} + (\mathbf{w}^T \mathbf{w}) \hat{\gamma} = \mathbf{w}^T \mathbf{y}. \quad (6.27)$$

Thus, $\hat{\gamma}$ satisfies

$$\begin{aligned} [\mathbf{w}^T \mathbf{w} - \mathbf{w}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{w}] \hat{\gamma} &= \mathbf{w}^T \mathbf{y} - \mathbf{w}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \mathbf{w}^T [\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \mathbf{y} \end{aligned} \quad (6.28)$$

so that

$$\hat{\gamma} = \frac{\mathbf{w}^T (\mathbf{I} - \mathbf{H}) \mathbf{y}}{\mathbf{w}^T (\mathbf{I} - \mathbf{H}) \mathbf{w}} = \frac{\langle \mathbf{w}, (\mathbf{I} - \mathbf{H}) \mathbf{y} \rangle}{\langle \mathbf{w}, (\mathbf{I} - \mathbf{H}) \mathbf{w} \rangle} = \frac{\langle (\mathbf{I} - \mathbf{H}) \mathbf{w}, (\mathbf{I} - \mathbf{H}) \mathbf{y} \rangle}{\langle (\mathbf{I} - \mathbf{H}) \mathbf{w}, (\mathbf{I} - \mathbf{H}) \mathbf{w} \rangle}. \quad (6.29)$$

But $(\mathbf{I} - \mathbf{H}) \mathbf{y} = \mathbf{y} - \hat{\mathbf{y}} = \hat{\boldsymbol{\varepsilon}}$, so that

$$\hat{\gamma} = \frac{\langle \hat{\boldsymbol{\varepsilon}}, (\mathbf{I} - \mathbf{H}) \mathbf{w} \rangle}{\langle (\mathbf{I} - \mathbf{H}) \mathbf{w}, (\mathbf{I} - \mathbf{H}) \mathbf{w} \rangle}. \quad (6.30)$$

From (6.29) and (6.30) we see that $\hat{\gamma}$ is the slope of the least squares line obtained by regressing $\hat{\varepsilon}$ on $\mathbf{w}_{res} = (\mathbf{I} - \mathbf{H})\mathbf{w}$, the residuals obtained by regressing \mathbf{w} on the columns of \mathbf{X} .

Now reversing the argument, suppose that \mathbf{w} is a column of the original model, say \mathbf{x}_j . Then, using (6.22) with $\mathbf{w} = \mathbf{x}_j$ and $\mathbf{X} = \mathbf{X}_{(-j)}$, where $\mathbf{X}_{(-j)}$ is the design matrix with \mathbf{x}_j deleted, then the least squares estimate $\hat{\beta}_j$ of β_j is given by

$$\hat{\beta}_j = \frac{\langle \hat{\varepsilon}_{(-j)}, \mathbf{x}_{j,res} \rangle}{\langle \mathbf{x}_{j,res}, \mathbf{x}_{j,res} \rangle} \quad (6.31)$$

and again this is the slope of the regression line obtained by plotting the residuals from the model without \mathbf{x}_j against $\mathbf{x}_{j,res}$, which is \mathbf{x}_j with the effect of $\mathbf{x}_i, i \neq j$ removed. This argument suggests plotting the residuals $\hat{\varepsilon}_{(-j)}$ against $\mathbf{x}_{j,res}, 1 \leq j \leq m$. From (6.31), it follows that if the model is true, then this plot should display a scatter of points about a line through the origin with slope $\hat{\beta}_j$. As before, deviations from linearity in \mathbf{x}_j , show up in a plot differing markedly from linearity. In essence, these plots play the role of scatter plots for SLR.

In addition, it can be shown that high leverage points in the response can be found in the extreme values of $\mathbf{x}_{j,res}$ [87]. Another interesting consequence of the formula (6.30) for $\hat{\gamma}$ is that it shows that multiple regression can be considered as a sequence of simple linear regressions by successively regressing an additional variable using the residuals from fitting the previously variables.

6.4 PRESS Residuals

So far in our discussion of residuals, we have dealt with residuals which arise from model fitting. However, if we are interested in using the model for prediction in addition to explanation, it would be useful to have a way of quantifying how well the model predicts in addition to how well it fits the observed data. One could use confidence intervals and prediction intervals, but unless we know exactly which points we wish to use, this approach does not seem to provide a convenient method for practical application. Perhaps the simplest approach to measuring the predictive power of the model would be to examine the residuals obtained by predicting values at new points. This is a “Catch-22” since we generally have no data at these points.

A convenient proxy for measuring the error at a new point is to use the data itself for this purpose. A reasonable approach, as indicated in Chapter 3, would be to omit one observation from the data, fit the model without this case and compare what the model predicts there, compared to the observation. If we omit the i -th observation, let $\hat{y}_{(-i)}$ be the predicted value from the model with the i -th observation removed. Then the value

$$\hat{\varepsilon}_{(-i)} = y_i - \hat{y}_{(-i)}, \quad 1 \leq i \leq n \quad (6.32)$$

is called the i -th *PRESS residual*. (PRESS is short for *prediction error sum of squares*.) The sum of squares

$$PRESS = \sum_{i=1}^n \hat{\varepsilon}_{(-i)}^2 \quad (6.33)$$

is a useful measure of predictive accuracy. Before discussing the use of PRESS as a diagnostic for model validity, it is interesting to consider the computation of $\hat{\varepsilon}_{(-i)}, 1 \leq$

$i \leq n$. If n is large, this looks like a daunting amount of computation, even with modern computers, since it appears that one reads to fit a new model to compute each $\hat{\varepsilon}_{(-i)}$. Remarkably, this turns out to be unnecessary. As we will see, it can be shown that

$$\hat{\varepsilon}_{(-i)} = \frac{\hat{\varepsilon}_i}{(1 - h_{ii})} \quad (6.34)$$

where $\hat{\varepsilon}_i$ is the residual from fitting the model to all of the data. Hence, all of these residuals can be obtained from factors obtained entirely from the original fit. Because these *deletion statistics* play an important role in modern regression analysis, we make a brief digression to establish (6.34) and some related results.

6.4.1 Deletion Statistics

We will refer to any of the regression statistics obtained by omitting one or more observations as *deletion statistics*. As we show, the trick to obtaining efficient formulas for these is to apply the Sherman-Woodbury-Morrison theorem in Section 4.4 to the design matrix \mathbf{X} with one row removed.

Let $\mathbf{X}_{(-i)}$ denote the design matrix resulting from deleting the i -th row, then the basic result we require is a formula relating $[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}]^{-1}$ and $(\mathbf{X}^T \mathbf{X})^{-1}$.

Theorem 6.1 *Let \mathbf{x}_i denote the i -th row of \mathbf{X} , then, if $h_{ii} \neq 1$*

$$[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}]^{-1} = \mathbf{X}^T \mathbf{X}^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}} \quad (6.35)$$

where h_{ii} is the i -th diagonal element of \mathbf{H} .

Before proving Theorem 6.1 we make the following crucial observation which relates $\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}$ to $\mathbf{X}^T \mathbf{X}$; i.e.,

$$\begin{array}{ccc} \mathbf{X}^T \mathbf{X} & = & \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} + \mathbf{x}_i^T \mathbf{x}_i \\ (m+1) \times (m+1) & & (m+1) \times (m+1) \quad (m+1) \times (m+1) \end{array} \quad (6.36)$$

For notational convenience, we assume that $i = n$ which can always be achieved by permuting the rows of \mathbf{X} . Then the ij -th element of $\mathbf{X}^T \mathbf{X}$ is given by

$$(\mathbf{X}^T \mathbf{X})_{ij} = \sum_{k=1}^n x_{ki} x_{kj} = \sum_{k=1}^{n-1} x_{ki} x_{kj} + x_{ni} x_{nj}. \quad (6.37)$$

Now the ij -th element of $\mathbf{x}_n^T \mathbf{x}_n$ is $x_{ni} x_{nj}$ and $\sum_{k=1}^{n-1} x_{ki} x_{kj}$ is the ij -th element of $\mathbf{X}_{(-n)}^T \mathbf{X}_{(-n)}$. Hence, (6.36) follows.

Proof of Theorem 6.1. Letting $\mathbf{A} = \mathbf{X}^T \mathbf{X}$, $\mathbf{B} = \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}$ and $\mathbf{z} = \mathbf{x}_i^T$ in Theorem 4.5 gives

$$[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}]^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1}}{1 - \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T}. \quad (6.38)$$

But $h_{ii} = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T$ so that (6.35) follows. ■

Theorem 6.2 Let $\hat{\varepsilon}_{(-i)}$ be the i -th PRESS residual, then

$$\hat{\varepsilon}_{(-i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}, \quad 1 \leq i \leq n. \quad (6.39)$$

Proof. Let $\hat{\boldsymbol{\beta}}_{(-i)}$ be the coefficient vector of parameter estimates omitting the i -th observation. Then,

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right]^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \quad (6.40)$$

where $\mathbf{y}_{(-i)}$ is the observation vector \mathbf{y} with y_i omitted. Then,

$$\hat{y}_{(-i)} = \langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}_{(-i)} \rangle = \left\langle \mathbf{x}_i, \left[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right]^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \right\rangle. \quad (6.41)$$

Using (6.38),

$$\begin{aligned} & \left\langle \mathbf{x}_i, \left[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right]^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \right\rangle \\ &= \left\langle \mathbf{x}_i, \left[(\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}} \right] \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \right\rangle \\ &= \left\langle \mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \right\rangle + \frac{\left\langle \mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \right\rangle}{1 - h_{ii}} \\ &= S_1 + S_2. \end{aligned} \quad (6.42)$$

Now using the fact that $\mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} = \mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i^T$,

$$\begin{aligned} S_1 &= \left\langle \mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i^T) \right\rangle \\ &= \left\langle \mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \right\rangle - y_i \left\langle \mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \right\rangle \\ &= \left\langle \mathbf{x}_i, \hat{\boldsymbol{\beta}} \right\rangle - h_{ii} y_i \\ &= \hat{y}_i - h_{ii} y_i. \end{aligned} \quad (6.43)$$

Similarly, the numerator of S_2 equals

$$\begin{aligned} & \left\langle \underbrace{\mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T}_{h_{ii}} \underbrace{\mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}}_{\hat{y}_i} \right\rangle - y_i \left\langle \underbrace{\mathbf{x}_i, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T}_{h_{ii}} \underbrace{\mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T}_{h_{ii}} \right\rangle \\ &= h_{ii} \hat{y}_i - h_{ii}^2 y_i. \end{aligned} \quad (6.44)$$

Thus,

$$\hat{y}_{(-i)} = \hat{y}_i - h_{ii} y_i + \frac{h_{ii} \hat{y}_i - h_{ii}^2 y_i}{1 - h_{ii}}. \quad (6.45)$$

A little elementary algebra then shows that

$$y_i - \hat{y}_{(-i)} = \frac{y_i - \hat{y}_i}{1 - h_{ii}} \quad (6.46)$$

which is (6.39). ■

For future reference we need analogous formulas for $\hat{\beta} - \hat{\beta}_{(-i)}$ and $SSE_{(-i)}$.

Theorem 6.3 (i) *As above, let $\hat{\beta}_{(-i)}$ be the least squares estimate of β with the i -th observation deleted, then,*

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_i}{1 - h_{ii}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_{(-i)}. \quad (6.47)$$

(ii) *If $SSE_{(-i)} = \sum_{j=1}^n [y_{(-i)} - \mathbf{x}_j^T \hat{\beta}_{(-i)}]^2$ is the error sum of squares for the fitted model with the i -th observation deleted, then,*

$$SSE_{(-i)} = \sum_{j=1}^n \hat{\epsilon}_j^2 - \frac{\hat{\epsilon}_i^2}{1 - h_{ii}}. \quad (6.48)$$

Proof. (i) As in Theorem 6.2

$$\begin{aligned} \hat{\beta}_{(-i)} &= \left(\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right)^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i^T) + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i^T)}{1 - h_{ii}} \\ &= S_1 + S_2. \end{aligned} \quad (6.49)$$

As before,

$$S_1 = \hat{\beta} - y_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{x}_i^T \quad (6.50)$$

and the numerator of S_2 is

$$\begin{aligned} &(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} - y_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \underbrace{\mathbf{x}_i \hat{\beta}}_{=\hat{y}_i} - \hat{y}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T h_{ii}. \end{aligned} \quad (6.51)$$

Thus,

$$\begin{aligned} \hat{\beta}_{(-i)} - \hat{\beta} &= -y_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{y}_i - h_{ii} y_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T}{1 - h_{ii}} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \left(-y_i + \frac{\hat{y}_i - h_{ii} y_i}{1 - h_{ii}} \right) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \left(\frac{-y_i + h_{ii} y_i + \hat{y}_i - h_{ii} y_i}{1 - h_{ii}} \right) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \left(\frac{\hat{y}_i - y_i}{1 - h_{ii}} \right). \end{aligned} \quad (6.52)$$

Thus, $\hat{\beta}_{(-i)} - \hat{\beta}_i = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_{(-i)}$ and changing signs gives (6.47).

(ii) By definition,

$$\begin{aligned} SSE_{(-i)} &= \left\langle \mathbf{y}_{(-i)} - \mathbf{x}_{(-i)} \hat{\beta}_{(-i)}, \mathbf{y}_{(-i)} - \mathbf{x}_{(-i)} \hat{\beta}_{(-i)} \right\rangle \\ &= \sum_{j=1, j \neq i}^n \left[y_j - \mathbf{x}_j \hat{\beta}_{(-i)} \right]^2 \\ &= \sum_{j=1}^n \left[y_j - \mathbf{x}_j \hat{\beta}_{(-i)} \right]^2 - \left[y_i - \mathbf{x}_i \hat{\beta}_{(-i)} \right]^2. \end{aligned} \quad (6.53)$$

From (i)

$$\hat{\beta}_{(-i)} = \hat{\beta} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_i}{1 - h_{ii}} \quad (6.54)$$

and using this in (6.53) it becomes

$$\sum_{j=1}^n \left[y_i - \mathbf{x}_j \hat{\beta} + \frac{\mathbf{x}_j (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_i}{1 - h_{ii}} \right]^2 - \left[y_i - \mathbf{x}_i \hat{\beta} + \frac{\mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_i}{1 - h_{ii}} \right]^2. \quad (6.55)$$

Using the fact that $h_{ij} = \mathbf{x}_j (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T$, this simplifies to

$$\sum_{j=1}^n \left(\hat{\epsilon}_j + \frac{h_{ij} \hat{\epsilon}_i}{1 - h_{ii}} \right)^2 - \left(\hat{\epsilon}_i + \frac{h_{ii} \hat{\epsilon}_i}{1 - h_{ii}} \right)^2 = \sum_{j=1}^n \left(\hat{\epsilon}_j - \frac{h_{ij} \hat{\epsilon}_i}{1 - h_{ii}} \right)^2 - \frac{\hat{\epsilon}_i^2}{(1 - h_{ii})^2}. \quad (6.56)$$

Expanding the sum in (6.56) gives

$$\sum_{j=1}^n \hat{\epsilon}_j^2 + \frac{2\hat{\epsilon}_i}{1 - h_{ii}} \sum_{j=1}^n h_{ij} \hat{\epsilon}_j + \frac{\hat{\epsilon}_i^2}{(1 - h_{ii})^2} \sum_{j=1}^n h_{ij}^2 - \frac{\hat{\epsilon}_i^2}{(1 - h_{ii})^2}. \quad (6.57)$$

From (6.11) $\sum_{j=1}^n h_{ij}^2 = h_{ii}$ and using the fact that $\mathbf{H}\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{H}\hat{\mathbf{e}} = 0$ (because $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \Rightarrow \mathbf{H}\hat{\mathbf{y}} = \mathbf{H}^2\mathbf{y} = \mathbf{H}\mathbf{y}$) $\sum_{j=1}^n h_{ij} \hat{\epsilon}_j = 0$. Using these in (6.57) gives

$$\begin{aligned} SSE_{(-i)} &= \sum_{j=1}^n \hat{\epsilon}_j^2 + \frac{\hat{\epsilon}_i^2 h_{ii}}{(1 - h_{ii})^2} - \frac{\hat{\epsilon}_i^2}{(1 - h_{ii})^2} \\ &= \sum_{j=1}^n \hat{\epsilon}_j^2 - \frac{\hat{\epsilon}_i^2}{1 - h_{ii}}, \end{aligned} \quad (6.58)$$

as required. ■

Corollary 6.1 For a model (5.1) with $m + 1$ parameters

$$E[SSE_{(-i)}] = (n - m - 2) \sigma^2, \quad (6.59)$$

so that

$$\hat{\sigma}_{(-i)}^2 = \frac{SSE_{(-i)}}{n - m - 2} \quad (6.60)$$

is an unbiased estimator of σ^2 . Thus

$$\hat{\sigma}_{(-i)}^2 = \frac{(1 - h_{ii})(n - m - 1)s^2 - \hat{\varepsilon}_i^2}{(1 - h_{ii})(n - m - 2)} \quad (6.61)$$

where s^2 is the SSE for the full model.

Proof. We prove (6.59), (6.60) and (6.61) are obvious consequences of (6.59) and (6.60). Since $E(\hat{\varepsilon}_i^2) = \text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, $1 \leq i \leq n$, it follows from (6.48) that

$$\begin{aligned} E[SSE_{(-i)}] &= \sum_{j=1}^n (1 - h_{jj})\sigma^2 - \sigma^2 \\ &= \sigma^2(n - 1) - \sigma^2 \sum_{j=1}^n h_{jj}. \end{aligned} \quad (6.62)$$

But $\sum_{j=1}^n h_{jj} = \text{tr}(\mathbf{H}) = m + 1$ as follows from the argument in Theorem 5.4. Thus, $E[SSE_{(-i)}] = (n - m - 2)\sigma^2$, as required. ■

We make several comments concerning the results in Theorems 6.1-6.3. Since it can be shown that $SSE_{(-i)}$ and $\hat{\varepsilon}_i$ are independent random variables and that $SSE_{(-i)}$ is a $\chi^2(n - m - 2)$ random variable (see Section 2.8), and $\hat{\varepsilon}_i/\sqrt{1 - h_{ii}}$ is $N(0, 1)$, then,

$$\hat{t}_i^* = \frac{\hat{\varepsilon}_i/\sqrt{1 - h_{ii}}}{\hat{\sigma}_{(-i)}} \quad (6.63)$$

has a t -distribution with $n - m - 2$ degrees of freedom. Of course \hat{t}_i^* is the same as RSTUDENT introduced in Section 6.3. As noted there, because the distribution of \hat{t}_i^* is known, one can use RSTUDENT to conduct tests of hypotheses concerning a particular observation being an outlier.

There is also an interesting relation between the PRESS residual and RSTUDENT. From (6.39) $E(\hat{\varepsilon}_{(-i)}) = 0$ and $\text{Var}(\hat{\varepsilon}_{(-i)}) = \sigma^2/(1 - h_{ii})$. Hence, if we define a standardized PRESS residual

$$\frac{\hat{\varepsilon}_{(-i)}}{\sigma(\hat{\varepsilon}_{(-i)})} = \frac{\hat{\varepsilon}_i/(1 - h_{ii})}{\sigma/\sqrt{1 - h_{ii}}} = \frac{\hat{\varepsilon}_i}{\sigma/\sqrt{1 - h_{ii}}} = \hat{r}_i. \quad (6.64)$$

A similar calculation shows that using $\hat{\sigma}_{(-i)}^2$ to estimate σ^2 , the studentized PRESS residual is given by $\hat{\varepsilon}_i/\hat{\sigma}_{(-i)}\sqrt{1 - h_{ii}} = \text{RSTUDENT}$. Hence, plots of studentized PRESS residuals are the same as plots of RSTUDENT.

Further, we note that formulas (6.63) and (6.64) have interesting diagnostic information. From (6.39) we see that if the i -th observation point has high leverage, then $\hat{\varepsilon}_{(-i)}$ will generally be much larger than $\hat{\varepsilon}_i$. Since high leverage points are fitted well, as measured by $\hat{\varepsilon}_{(-i)}$ they may predict poorly. Again, this is another manifestation of the fit/prediction dilemma. How to account for both is a continuing conundrum in statistical analysis.

This same phenomenon shows up in (6.47) for $\hat{\beta}_i - \hat{\beta}_{(-i)}$, which shows that the influence of the i -th observation depends again on $\hat{\varepsilon}_{(-i)}$. So, it can be “small” if the fit is good but large if h_{ii} is large. Here, again, fit and leverage act in opposite directions. How to reconcile this is considered next.

6.4.2 Influence Diagnostics

Until recently, the emphasis in multiple regression analysis has been to determine the appropriate variables to include in the model (5.1). From a data point of view, the emphasis was placed on the relations among the columns of \mathbf{X} . This has already been well illustrated in Chapter 5 and will be further elaborated on in Chapters 8 and 9.

Since we only have a finite sample of (\mathbf{X}, \mathbf{y}) values to work with, even if the model is perfect, it might be possible to obtain different conclusions if a different sample $(\mathbf{X}', \mathbf{y}')$ was obtained. Unfortunately in most situations we will only have the original data (\mathbf{X}, \mathbf{y}) to work with and it then becomes important to consider how this particular sample affects our conclusions. Hence, this leads us to focus on the effects that the rows of \mathbf{X} have on the estimated model. Of course if we had multiple samples we could compare the effect of different observations on the estimated model. Since generally this cannot be done, we must look for a different approach similar to the use of PRESS residuals. That is, we consider omitting variables one or more at a time to see how $\hat{\beta}$ and $\hat{\mathbf{y}}$ change. Clearly, even if the number of observations is moderate, this can lead to the examination of large amount of data (if $n = 10$, there could be 1023 different regressions). However, we can compute the effect of deleting observations without having to calculate any additional regressions, as we have already done for computing the PRESS residuals in the previous section. These were required for the computation of $\hat{\beta} - \hat{\beta}_{(-i)}$. As we shall see, these formulas are useful for investigating the effect of individual data points on the estimation of β and the fit $\hat{\mathbf{y}}$. Although formulas can be obtained for the deletion of subsets containing more than one observation, we will only consider single point influence statistics here.

Before proceeding with the more formal analysis, we discuss the main reason for doing this. Since $\hat{\beta}$ and $\hat{\mathbf{y}}$ both depend on (\mathbf{X}, \mathbf{y}) it is important to know whether or not all points are contributing equally to the estimation or whether there are points which are unduly affecting it. For example, points with large leverage h_{ii} indicate that the i -th observation overly influences the fit, while points with large residuals suggest possible model inadequacy. It is important to know if these discrepant points should be used, set aside or reexamined further as to their occurrence. As we have already observed, high leverage can mask residuals when using \hat{r}_i or \hat{t}_i , examination of these alone may not be sufficient to detect unusual influential points.

As we shall see, examination of the statistic $\hat{\beta} - \hat{\beta}_{(-i)}$ and $\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)}$ enable us to combine these two competing factors to mathematically assess the influence of the i -th observation $(\mathbf{x}_i, \mathbf{y}_i)$.

6.4.3 Influence on $\hat{\beta}$, $\hat{\mathbf{y}}_i$

Since $\hat{\beta} - \hat{\beta}_{(-i)}$ clearly measures the effect of deleting the i -th observation on $\hat{\beta}$, this should form the basis of influence diagnostics of \mathbf{x}_i or $\hat{\beta}$. First, if the effect on an individual coefficient is desired, one can then consider the j -th component of

$$\left(\hat{\beta} - \hat{\beta}_{(-i)}\right)_j = \hat{\beta}_j - \hat{\beta}_{(-i)j}. \quad (6.65)$$

Using $\hat{\beta} - \hat{\beta}_{(-i)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\epsilon}_i / (1 - h_{ii})$ as follows from (6.65) we get

$$\hat{\beta}_j - \hat{\beta}_{(-i)j} = \frac{r_{ji} \hat{\epsilon}_i}{(1 - h_{ii})} \quad (6.66)$$

where r_{ji} is ji -th element of $\mathbf{R} \equiv (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

One generally considers the i -th observation influential on the j -th coefficient β_j if the quantity in (6.66) is considered to be “large.” Since $\hat{\beta}_j$ is a random variable, large should be measured relative to the standard error σ_j of $\hat{\beta}_j$, which is $\sigma\sqrt{\delta_j}$ where δ_j is the j -th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$. If we estimate $\sigma\sqrt{\delta_j}$ by $\hat{\sigma}_{(-i)}\sqrt{\delta_j}$ then dividing (6.66) by this leads to the influence statistic $DFBETAS_{j,i}$

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{(-i)j}}{\hat{\sigma}_{(-i)}\sqrt{\delta_j}} = \frac{r_{ji}\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{\delta_j}(1 - h_{ii})} \quad (6.67)$$

which gives the number of standard errors that the coefficient changes if the i -th observation were set aside. A large value (in magnitude) of $DFBETAS_{j,i}$ indicates that the i -th observation has a sizable impact on the j -th regression coefficient. The sign of $DFBETAS_{j,i}$ may also be meaningful. For example, if $DFBETAS_{j,i}$ in (6.67) is negative and relatively large in magnitude, it is likely that the negative coefficient can be attributed to the i -th observation.

Furthermore,

$$\mathbf{R}\mathbf{R}^T \equiv (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{R}^T \mathbf{R} = \boldsymbol{\delta} \text{ (say)}. \quad (6.68)$$

Therefore, $\delta_j = \mathbf{r}_j^T \mathbf{r}_j$ where \mathbf{r}_j^T denotes the j -th row of \mathbf{R} . Using this (6.67) can be written

$$DFBETAS_{j,i} = \frac{r_{ji}}{\sqrt{\mathbf{r}_j^T \mathbf{r}_j}} \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}(1 - h_{ii})} = \frac{r_{ji}}{\sqrt{\mathbf{r}_j^T \mathbf{r}_j}} \frac{\hat{t}_i}{\sqrt{1 - h_{ii}}} \quad (6.69)$$

where \hat{t}_i is the RSTUDENT. Again, $DFBETAS_{j,i}$ measures both leverage h_{ii} and the effect (errors) of a large residual.

In [8] Belsley, Kuh and Welsch suggest a cut-off value of $2/\sqrt{n}$ for $DFBETAS_{j,i}$. If this quantity is exceeded then observation i is considered to be influential on the estimation of β_j . Since $DFBETAS_{j,i}$ is the ji -th element of an $(m+1) \times n$ matrix this can lead to a substantial amount of output to examine even for moderate value of (m, n) .

To mitigate this problem it is somewhat easier to examine the effect of the i -th observation by considering its effect on the whole coefficient vector $\hat{\boldsymbol{\beta}}$. This can be done by using some norm of the vector $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}$ as suggested by Cook in [18, 19]. The general form of this statistic is

$$D_i = \frac{\langle \hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_{(-i)}, \mathbf{M}(\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_{(-i)}) \rangle}{(m+1)c} \quad (6.70)$$

where \mathbf{M} is a given positive definite matrix and c is a normalizing constant. The most popular choice of \mathbf{M} is $(\mathbf{X}^T \mathbf{X})$ and $c = s^2$, the usual unbiased estimate of variance. In this case

$$D_i = \frac{\langle \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}, (\mathbf{X}^T \mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}) \rangle}{(m+1)s^2}, \quad (6.71)$$

which is called *Cook's Distance*. In computer output that is usually referred to as *Cook's*

D. For computational purposes using (6.71)

$$\begin{aligned}
 \left\langle \hat{\beta} - \hat{\beta}_{(-i)}, (\mathbf{X}^T \mathbf{X}) (\hat{\beta} - \hat{\beta}_{(-i)}) \right\rangle &= \left\langle \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \frac{\mathbf{x}_i^T \hat{\varepsilon}_i}{1 - h_{ii}}, (\mathbf{X}^T \mathbf{X}) \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \hat{\varepsilon}_i}{1 - h_{ii}} \right\rangle \\
 &= \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2 \left\langle (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T, \mathbf{x}_i^T \right\rangle \\
 &= \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2 \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \\
 &= \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2 h_{ii}.
 \end{aligned} \tag{6.72}$$

Hence

$$D_i = \frac{\hat{\varepsilon}_i^2 h_{ii}}{(1 - h_{ii})^2 (m + 1) \hat{\sigma}_{(-i)}^2}. \tag{6.73}$$

Recalling that $\hat{t}_i = \hat{\varepsilon}_i / \hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}$

$$D_i = \frac{\hat{t}_i^2}{(m + 1)} \left(\frac{h_{ii}}{1 - h_{ii}} \right). \tag{6.74}$$

Hence, no new calculations need to be done, since one generally will have computed \hat{t}_i and h_{ii} in the course of the analysis.

Also, note that D_i combines the effect of fit through the residual \hat{t}_i and leverage through h_{ii} . Since $h_{ii} / (1 - h_{ii})$ is an increasing function of h_{ii} , D_i can be large if either \hat{t}_i and/or h_{ii} is large (i.e., close to one). Again, this points out the need to isolate those points with either large residuals (outliers) or high leverage.

As before, the question arises for the need to have some cut-off value for which D_i is to be considered large.

By comparing the form of D_i to that for the F -statistic used to compute the joint confidence region (5.282) for β , it is suggested that values of $D_i > 1$ are to be considered large. Hence, we consider every point with $D_i > 1$ to be influential on the estimate $\hat{\beta}$ and should be set aside for further examination.

Finally we note that D_i can be written in the alternative form

$$D_i = \frac{\langle \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)}, \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)} \rangle}{(m + 1) s^2} \tag{6.75}$$

which follows easily from (6.71). We leave the details to the reader. Hence, D_i may also be considered as a statistic to measure the influence of the i -th observation on the overall fit.

One may also consider the influence of the i -th observation on the individual fitted values \hat{y}_i , $1 \leq i \leq n$. For this we define

$$DFFITs_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{\hat{\sigma}_{(-i)} \sqrt{h_{ii}}} \tag{6.76}$$

where the denominator is an estimate of $\sigma \left(\hat{Y}_i \right)$.

Using an argument similar to that for D_i we find that

$$DFFITS_i = \hat{t}_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}. \quad (6.77)$$

Again, we see that both fit and leverage determine the influence of the i -th observation on the fit. If either \hat{t}_i is large and/or h_{ii} is large (close to one) then $DFFITS_i$ will be large. However, a large residual \hat{t}_i (outlier) can be masked if $h_{ii} \simeq 1$.

In [8] Belsley, Kuh and Welsch suggest a cut-off value of $2\sqrt{(m+1)/n}$ for $DFFITS_i$ to be large. Other influence statistics are discussed in [8, 87, 20]. We now examine a couple of examples to see how one can use these diagnostics in practice.

Example 6.4 (Drink delivery data) Consider again the drink delivery data of Table 5.3. Table 6.5 shows the leverage h_{ii} , Cook's D , and DFFITS values, illustrating the influence of all 25 observations without removing any of them.

Table 6.5 h_{ii} , D_i and DFFITS for the data in Table 5.3

No.	h_{ii}	D_i	DFFITS	No.	h_{ii}	D_i	DFFITS
1	0.10180	0.10009	-0.5709	14	0.07824	0.00329	0.0974
2	0.07070	0.00338	0.0987	15	0.04111	0.00063	0.0426
3	0.09874	0.00001	-0.0052	16	0.16594	0.00329	-0.0972
4	0.08538	0.07765	0.5008	17	0.05943	0.00040	0.0339
5	0.07501	0.00054	-0.0395	18	0.09626	0.04398	0.3653
6	0.04287	0.00012	-0.0188	19	0.09645	0.01192	0.1862
7	0.08180	0.00217	0.0790	20	0.10169	0.13244	-0.6718
8	0.06373	0.00305	0.0938	21	0.16528	0.05086	-0.3885
9	0.49829	3.41932	4.2961	22	0.39158	0.45105	-1.1950
10	0.19630	0.05385	0.3987	23	0.04126	0.02990	-0.3075
11	0.08613	0.01620	0.2180	24	0.12061	0.10232	-0.5711
12	0.11366	0.00160	-0.0677	25	0.06664	0.00011	-0.0176
13	0.06113	0.00229	0.0813				

First, in order to consider the leverages, we calculate the cut-off leverage point, which is $2(m+1)/n = 2(3)/25 = 0.24$. Based on this criterion, observations 9 and 22 are high leverage points. Since the hat diagonal matrix provides a measure of standardized distance from the point \mathbf{x}_i to $\bar{\mathbf{x}}$ (and the reader should note that h_{ii} does not involve the y 's), these two observations exert undue influence on at least one regression coefficient as well as other performance criteria.

Now consider Cook's Distance, which considers both the location of the point in the x -space and the response variable in measuring influence. From Table 6.5, $D_9 = 3.42$ shows the largest value, which indicates that the observation 9 is definitely influential.

From (6.77), the cut-off value for $DFFITS_i$ for the drink delivery data is $2\sqrt{(2+1)/25} = 0.6928$. Inspecting Table 6.5 we notice that observations 9 and 22 have values of DFFITS that exceed the value, and also $|DFFITS_{20}|$ is close to the cut-off value.

The $DFBETAS_{j,i}$ ($j = 0, 1, 2$) values for all 25 observations were calculated and given in Table 6.6. The cut-off value is $2/\sqrt{25} = 0.4$. We then instantly notice that observation numbers 9 and 22 have large effects on all three regression coefficients as well as the quality of fit. As we see, observation 9 has a very large effect on the intercept and relatively smaller effects on $\hat{\beta}_1$ and $\hat{\beta}_2$, while the observation 22 has its largest effect

on $\hat{\beta}_1$. Besides these, several other observations produce effects on the coefficients that are close to the formal cut-off, including observations 1 (on $\hat{\beta}_1$ and $\hat{\beta}_2$), 4 (on $\hat{\beta}_0$), and 24 (on $\hat{\beta}_1$ and $\hat{\beta}_2$). These points produce relatively small changes in comparison to the observation 9.

From the diagnostic point of view, clearly the influence of observation 9 is evident, since its deletion results in a displacement of every regression coefficient by at least 0.9 standard deviations. The impact of observation 22 is much smaller. Furthermore, deleting observation 9 displaces the predicted response by over four standard deviations.

Table 6.6 DFBETAS_{*j,i*} values for the data in Table 5.3

No. <i>i</i>	Intercept (<i>j</i> = 0)	Cases (<i>j</i> = 1)	Distance (<i>j</i> = 2)	No. <i>i</i>	Intercept (<i>j</i> = 0)	Cases (<i>j</i> = 1)	Distance (<i>j</i> = 2)
1	-0.1873	0.4113	-0.4349	14	0.0495	-0.0671	0.0618
2	0.0898	-0.0478	0.0144	15	0.0223	-0.0048	0.0068
3	-0.0035	0.0039	-0.0028	16	-0.0027	0.0644	-0.0842
4	0.4520	0.0883	-0.2734	17	0.0289	0.0065	-0.0157
5	-0.0317	-0.0133	0.0242	18	0.2486	0.1897	-0.2724
6	-0.0147	0.0018	0.0011	19	0.1726	0.0236	-0.0990
7	0.0781	-0.0223	-0.0110	20	0.1680	-0.2150	-0.0929
8	0.0712	0.0334	-0.0538	21	-0.1619	-0.2972	0.3364
9	-2.5757	0.9287	1.5076	22	0.3986	-1.0254	0.5731
10	0.1079	-0.3382	0.3413	23	-0.1599	0.0373	-0.0527
11	-0.0343	0.0925	-0.0027	24	-0.1197	0.4046	-0.4654
12	-0.0303	-0.0487	0.0540	25	-0.0168	0.0008	0.0056
13	0.0724	-0.0356	0.0113				

Example 6.5 (Housing data) Consider the housing price data. There are two regressors, and the data were used to illustrate the influence diagnostics. Table 6.7 shows residuals and some of those essential diagnostic statistics.

Table 6.7 Residuals and Diagnostic Statistics for Housing Data

No.	Res $\hat{\epsilon}_i$	T Res \hat{t}_i	h_{ii}	Cook's <i>D</i>	DFFITs
1	-572.1	-0.07582	0.13335	0.000321	-0.02974
2	8876.8	1.23772	0.10868	0.059620	0.43219
3	-9137.4	-1.36422	0.20219	0.146691	-0.68677
4	-3337.5	-0.44779	0.13963	0.011621	-0.1804
5	-4858.5	-0.73303	0.29917	0.079524	-0.4790
6	-15795.7	-2.94203	0.21654	0.486842	-1.5467
7	-4067.9	-0.57252	0.20924	0.030626	-0.2945
8	11025.1	1.79205	0.25621	0.311363	1.0518
9	4383.9	0.59898	0.15862	0.023819	0.2601
10	4227.7	0.56893	0.13536	0.017900	0.2251
11	4141.5	0.62072	0.29917	0.057784	0.4056
12	-4543.4	-0.64612	0.21935	0.041095	-0.3425
13	6891.1	1.23240	0.45877	0.411341	1.1346
14	-1068.6	-0.13753	0.08021	0.000599	-0.0406
15	3834.9	0.49962	0.08353	0.008090	0.1508

We first calculate the cut-off leverage point, which is $2(m+1)/n = 2(3)/15 = 0.4$. Data point 10 is closest to the cut-off leverage point. The cut-off value for DFFITS_i is $2\sqrt{(2+1)/15} = 0.8944$. Data points 6, 8 and 13 have larger values than the cut-off value. These are considered to be influential.

The $\text{DFBETAS}_{j,i}$ ($j = 0, 1, 2$) values for all 15 observations can also be calculated using either (6.67) or (6.69). We leave these to the reader for an exercise. The cut-off value is $2/\sqrt{15} = 0.5164$.

6.5 Transformations

As for simple linear regression, if examination of the model using the techniques of Chapter 5 and this (or others) indicates that modifications should be made, then various transformations may be required to provide a more adequate representation of the data. This can require transforming the data in \mathbf{y} or \mathbf{X} , adding or deleting variables and/or observations, changing the error structure and/or changing the method of estimation. In this section we focus on transforming the existing data and modifying least squares to account for violations in the basic assumption of the GLM.

In Chapter 8 we will consider further modeling issues concerning the addition of new variables while in Chapter 9, we will consider modifications to deal with multicollinearity. As there will be some overlap with our discussion Chapter 3 we will refer there for details as necessary.

6.5.1 Transformations in \mathbf{x}

If various diagnostics or theoretical considerations indicate that the dependence of \mathbf{y} on \mathbf{x} is not linear in one or more of the independent variables, then it may be appropriate to reformulate the model by transforming the variables in some fashion as indicated for SLR in Chapter 3. For example, suppose we suspect nonlinearity in the j -th variable x_j in (5.1), then there are a number of ways to proceed.

One can replace x_j by a variable $z_j = f(x_j)$ so the model (5.1) becomes

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j z_j + \dots + \beta_m x_m + \varepsilon. \quad (6.78)$$

If the functional form $z_j = f(x_j)$ is known, then (6.78) is a linear model and can be fit and analyzed as we have already discussed. If this transformation is appropriate over our initial assumption, then this should show up in improved fit statistics, such as R^2 , t values, increased F and reduced curvature of residual plots involving z_j over those involving x_j . Unfortunately, the functional form $z_j = f(x_j)$ is usually unknown, so an analysis generally has to include methods for doing this. A variety of exploratory methods can be found in [115]. A reasonable approach is to parametrize the unknown function in some fashion and then estimate these parameters along with β_i , $0 \leq i \leq m, i \neq j$. A typical parameterization is $z_j = x_j^\lambda$, where λ is an appropriate real number. If x_j is positive, then all values of λ are permissible while λ has to be restricted if x_j takes on negative values. A typical range of λ is $-2 \leq \lambda \leq 2$.

Since polynomials of suitable degree can be used to approximate fairly arbitrary

functions, one might consider looking for z_j as a polynomial in x_j , i.e.;

$$z_j = \sum_{k=1}^l r_k x_j^k \quad (6.79)$$

where $r_k, 1 \leq k \leq l$ have to be estimated. Using this in (6.78) the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_j \left(\sum_{k=1}^l r_k x_j^k \right) + \cdots + \beta_m x_m + \varepsilon. \quad (6.80)$$

Unfortunately this is not a linear model in the parameters $\beta_i, 0 \leq i \leq m, i \neq j$ and $r_k, 1 \leq k \leq l$. Since there are a number of pitfalls in using polynomials, we will reserve our treatment of this until the following Chapter. Other functional forms may be used as well, such as *trigonometric functions* and *piecewise polynomials (splines)*. Again we will take this up in Chapter 7. For now we concentrate on the power family.

6.5.2 The Box-Tidwell Method

One approach to estimating λ in $z_j = x_j^\lambda$ is to choose a range of λ , fit the model, for each value of λ and then choose the model which provides the “best” fit, say the one with the smallest *SSE* or largest R^2 or F . For a large range of λ , this may require a substantial amount of computation and/or analysis. In addition, because the behavior of test statistics as a function of λ is not known, this procedure could miss the best λ if it was not in the original parameter range. A more automatic procedure, analogous to the Box-Cox approach is to try to choose λ in some automatic fashion. We discuss one such procedure here; the *Box-Tidwell method* [12].

Suppose we assume that λ is not too different from $\lambda = 1$. Then expanding x^λ in a Taylor series about $\lambda = 1$ gives

$$x^\lambda \simeq x^1 + (\lambda - 1) \frac{d}{d\lambda} x^\lambda \Big|_{\lambda=1}. \quad (6.81)$$

From calculus,

$$\frac{d}{d\lambda} x^\lambda = x^\lambda \log x \quad (6.82)$$

so that $dx^\lambda/d\lambda|_{\lambda=1} = x \log x$.

Thus,

$$x^\lambda \simeq x + (\lambda - 1) x \log x \quad (6.83)$$

and using this approximation in (5.1) the assumed model is of the form

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_j [x_j + (\lambda - 1) x_j \log x_j] + \cdots + \beta_m x_m + \varepsilon \\ &= \beta_0 + \sum_{k=1}^m \beta_k x_k + \beta_j (\lambda - 1) x_j \log x_j + \varepsilon. \end{aligned} \quad (6.84)$$

Letting $\beta_{m+1}(\lambda) = \beta_j (\lambda - 1)$ (6.84) is a linear model in the variables $\beta_k, 0 \leq k \leq m + 1$. Since $\beta_{m+1} = (\lambda - 1) \beta_j$ we estimate (λ, β_j) in the following way. First, fit the original linear model and obtain the least squares estimate $\hat{\beta}_j$ of β_j . Then fit the enlarged

model with $x_{m+1} = x_j \log x_j$ by least squares and obtain the least squares estimate of $\beta_{m+1}, \hat{\beta}_{m+1}$. Using the relation $\hat{\beta}_{m+1} = (\hat{\lambda} - 1) \hat{\beta}_j$ gives

$$\hat{\lambda} = \left(\hat{\beta}_{m+1} / \hat{\beta}_j \right) + 1, \quad (6.85)$$

where $\hat{\beta}_{m+1}$ is least squares estimate of β_{m+1} . This allows one to test for the need for a transformation:

$$H_0 : \lambda = 1 \text{ against } H_1 : \lambda \neq 1 \quad (6.86)$$

by using the t test to test for $\beta_{m+1} = 0$. In addition, a $100 \times (1 - \alpha) \%$ CI for λ is given by

$$\left(\hat{\beta}_{m+1} + 1 \right) \pm t_{\alpha, n-m-2} \cdot s \quad (6.87)$$

where $s = \sqrt{SSE / (n - m - 2)}$.

In Atkinson's modification of this procedure, one applies Andrews' approach to the transformed data $z^{(\lambda)}$ in (6.93). Further details can be found in [20, 27].

Again, the new model (6.84) with $\hat{\lambda}$ given by (6.85) can be analyzed as before. As indicated in [20] it is useful to examine the added variable plot using the constructed variable $x_j \log x_j$ as a diagnostic for a transformation. A distinct trend in the plot suggests that $\lambda \neq 1$ in (6.84).

If the model (6.84) with $\hat{\lambda}$ appears inadequate, one can obtain successive estimates $\hat{\lambda}(l), l \geq 1$ by letting $\hat{\lambda}(0) = \hat{\lambda}$ in (6.85) and then getting a new estimate by expanding $x_j^{\hat{\lambda}}$ about $\lambda = \lambda(0)$ giving $x_j^{\hat{\lambda}} \simeq x_j^{\lambda(0)} + \{\lambda - \lambda(0)\} x_j^{\lambda(0)} \log x_j$ and substituting into (6.84) to give

$$Y = \beta_0 + \sum_{k=1, k \neq j}^m \beta_k x_k + \beta_j x_j^{\lambda(0)} + \beta_j [\lambda - \lambda(0)] x_j^{\lambda(0)} \log x_j + \varepsilon. \quad (6.88)$$

Then, fit the model (6.88) with and without the added variable $x_{m+1} = x_j^{\lambda(0)} \log x_j$. If $\hat{\beta}_j(1)$ and $\hat{\beta}_{m+1}(1)$ are the least squares estimate of β_j and β_{m+1} then

$$\hat{\lambda}(1) = 1 + \hat{\beta}_{m+1}(1) / \hat{\beta}_j(1). \quad (6.89)$$

Again, one can further iterate if the method is converging, or stop after a fixed number of iterations if not.

This method can be easily generalized to account for simultaneous transformations in two or more predictors and to consider transformations from $\lambda = \lambda_0$ rather than $\lambda = 1$. (Just use (6.88).)

Example 6.6 Consider the data set given in below. Suppose that we have a multiple regression problem with $m = 2$ predictors and we contemplate transforming one of them, say x_2 . We apply the Box-Tidwell method to these data and examine the need to transform x_2 by fitting

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

The overall F -statistic is 1621.74. Clearly, the model is significant at 1% and the results are $\hat{\beta}_2 = 0.32071$, its standard error is 0.02919, and t -value is 10.99.

Obs. No.	x_1	x_2	y
1	1	1	2.95
2	1	2	3.48
3	1	3	3.65
4	2	1	4.11
5	2	2	4.35
6	2	3	4.88
7	3	1	5.20
8	3	2	5.48
9	3	3	5.84
10	4	4	7.22
11	5	5	8.42

We now consider an enlarged model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 (\lambda - 1) x_2 \log x_2 + \varepsilon.$$

Then, the estimate of $\beta_2 (\lambda - 1)$ is -0.1593 , its standard error is 0.09812 , and the t -value is -1.62 . (To provide enough evidence of the need to transform the predictor, the t -value is compared to $t(v)$, for this example $v = 10$.)

Hence,

$$\hat{\lambda} = \frac{-0.1593}{0.32071} + 1 = 0.5033.$$

Rounding to a convenient multiple, we choose $\hat{\lambda} \simeq 1/2$ and take the square root of x_2 .

Thus, the fitted model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{x_2} + \varepsilon_i,$$

so the estimated regression surface is

$$\hat{y} = 0.984 + 1.07x_1 + 0.951\sqrt{x_2}.$$

6.5.3 Transformations of \mathbf{y}

As for SLR, transformation of \mathbf{y} in (5.1) is suggested by curvature in residual plots, in particular those of $\hat{\varepsilon}$ against $\hat{\mathbf{y}}$. In general, transformations in \mathbf{y} result from the following factors:

- (i) the relation between $Y_{\mathbf{x}}$ and \mathbf{X} is not linear;
- (ii) the variance of $Y_{\mathbf{x}}$ is not constant;
- (iii) the error model for $Y_{\mathbf{x}}$ is incorrect.

Often (ii) is a consequence of (iii). For example, in many situations the dependent variable $Y_{\mathbf{x}}$ is discrete, rather than continuous as implied by the assumption of normal errors. A typical, and increasingly important case is that of *logistic regression*. Here, $Y_{\mathbf{x}}$ may only take on the values 0 and 1, to record the success or failure of an experiment. In this case $Y_{\mathbf{x}}$ is a Bernoulli random variable, with $E(Y_{\mathbf{x}}) = p_{\mathbf{x}} = P\{Y_{\mathbf{x}} = 1\}$.

Since $\text{Var}(Y_{\mathbf{x}}) = p_{\mathbf{x}}(1 - p_{\mathbf{x}})$, the variance of $Y_{\mathbf{x}}$ will generally not be constant if $p_{\mathbf{x}}$ depends in a nontrivial way on the independent variable \mathbf{x} . Many other such situations occur in practice, such as count data, which typically follow a Poisson distribution [40] or in the analysis of survival times, when the error distribution may be exponential.

In this section we will consider transformations which deal with (i)-(ii), when data can be transformed to the normal model and in Chapter 7 we will consider some aspects of (iii).

6.5.4 Linearizable Transformations

As for SLR, a common approach to transforming Y , is to assume that there exists a transformation which linearizes the model. As a typical case, as in Chapter 3, is the assumption that

$$y = \exp \left(\sum_{j=0}^m \beta_j x_j \right) \quad (6.90)$$

so that

$$\log y = \sum_{j=0}^m \beta_j x_j. \quad (6.91)$$

Of course, as we pointed out in Section 3.10, taking logarithms in (6.90) will not necessarily transform the model to one with $N(0, \sigma^2)$ errors, unless the distribution of Y is lognormal. Otherwise, transforming Y transforms the model to one with a complicated error structure. In such cases, other estimation techniques, such as *nonlinear regression* may be needed to fit the original data.

If one refers to Table 3.24, any functional form given there can be used to linearize a nonlinear model but the same remarks concerning the error structure should be accounted for. If such a transform is suggested, one can fit the transformed model and then evaluate its appropriateness by the various plots and diagnostics discussed in this Chapter.

6.5.5 Box-Cox Transformations

As in Section 3.10, if a linearizing transformation is not immediately apparent, then one can use the Box-Cox method to determine a transformation of power type, if the data $y_i, 1 \leq i \leq n$, are positive. As there, we assume that

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log Y_i, & \lambda = 0, \end{cases} \quad (6.92)$$

is $N(0, \sigma^2)$. Then one can use MLE to estimate $\beta(\lambda)$, $\sigma(\lambda)$ and λ . Essentially repeating the calculations given in Section 3.10 this can be done by fixing λ and then regressing the modified data

$$z_i^{(\lambda)} = \frac{y_i^{(\lambda)}}{\bar{y}^{\lambda-1}} \quad (6.93)$$

where $\bar{y} = (\prod_{i=1}^n y_i)^{1/n}$ is the geometric mean of $y_i, 1 \leq i \leq n$. That is, we use the model

$$Z_i^{(\lambda)} = \sum_{j=1}^m \beta_j x_{ij} + \varepsilon_i \quad (6.94)$$

where $\varepsilon_i, 1 \leq i \leq n$ are $N(0, \sigma^2)$.

The MLE of $\sigma, \hat{\sigma}(\lambda)$ is given by

$$\hat{\sigma}(\lambda) = \sqrt{SSE(\lambda)/n}. \quad (6.95)$$

To obtain the MLE, $\hat{\lambda}$ of λ , one can proceed as for SLR by choosing a range of values of λ , often $-2 \leq \lambda \leq 2$ is a reasonable choice, and plotting $SSE(\lambda)$ against λ . The minimum value can then be estimated visually or by a more formal interpolation process.

6.5.6 Quick Estimates of λ

Because the Box-Cox method can be computationally intensive, a number of authors have suggested “quick transformations” to estimate λ . We discuss one here due to Andrews [3], which is analogous to the Box-Tidwell method. A modification of this method was given by Atkinson in [6].

As in the Box-Cox method we begin by trying to determine a value of λ such that $y^{(\lambda)}$ in (6.92) satisfies the conditions of the GLM. If $\lambda = 1$, then the data will not require transformation. To test this, we expand $y^{(\lambda)}$ in a Taylor series about $\lambda = 1$ giving

$$y^{(\lambda)} = y^1 + (\lambda - 1) \left. \frac{dy^{(\lambda)}}{d\lambda} \right|_{\lambda=1}. \quad (6.96)$$

From calculus,

$$\frac{d}{d\lambda} \left(\frac{y^\lambda - 1}{\lambda} \right) = \frac{y^\lambda \log y}{\lambda} - \frac{(y^\lambda - 1)}{\lambda^2} \quad (6.97)$$

so that

$$\left. \frac{d}{d\lambda} y^{(\lambda)} \right|_{\lambda=1} = y \log y - y + 1. \quad (6.98)$$

Hence, near $\lambda = 1$,

$$y^{(\lambda)} \simeq y - 1 + (\lambda - 1)(y \log y - y + 1). \quad (6.99)$$

Using (6.99) and (6.96) we get an approximate model

$$Y - 1 + (\lambda - 1)(Y \log Y - Y + 1) \simeq \beta_0 + \sum_{j=1}^m \beta_j x_j + \varepsilon \quad (6.100)$$

and rewriting this gives

$$Y \simeq 1 + \beta_0 + \sum_{j=1}^m \beta_j x_j + (1 - \lambda)(Y \log Y - Y + 1) + \varepsilon. \quad (6.101)$$

Letting

$$\beta'_0 = 1 + \beta_0, \beta_{m+1} = 1 - \lambda \quad \text{and} \quad x_{m+1} = \hat{y} \log \hat{y} - \hat{y} + 1 \quad (6.102)$$

where \hat{y} are the fitted values for the untransformed model ($\lambda = 1$) gives the further approximation

$$y \simeq \beta'_0 + \sum_{j=1}^m \beta_j x_j + \beta_{m+1} x_{m+1} + \varepsilon. \quad (6.103)$$

In this form, (6.103) is a standard linear model which can be fit by least squares. The estimate $\hat{\beta}_{m+1}$ can then be used to test for a need to transform the data. Since, we are considering the null hypothesis that $\lambda = 1$, this is equivalent to testing $\beta_{m+1} = 0$ in (6.103). This can be done using the usual t -test.

Using this test, if we conclude that $\beta_{m+1} \neq 0$, then we can transform the data using the estimate $\hat{\lambda}$ of λ given by

$$\hat{\lambda} = 1 - \hat{\beta}_{m+1} \quad (6.104)$$

and refit using least squares and noting any improvements.

To simplify matters, it was shown in [20] that the same t values for $\hat{\beta}_{m+1}$ are obtained if we use $x_{m+1} = \hat{y} \log \hat{y}$ rather than $x_{m+1} = \hat{y} \log \hat{y} - \hat{y} + 1$ in (6.103).

Example 6.7 Here we consider the need for a possible transformation of the housing price data in Example 5.12 and the birth weight data in Example 5.15.

For the housing prices we observed that observation 6 was an outlier and for the birth weight data there is a substantial amount of unexplained residual variation. Although the various statistics and plots suggest that the fitted models (5.98) and (5.103) are appropriate, we entertain the possibility that a transformation might improve the fits.

To examine this possibility we used the Andrews test for the added variable $x_3 = \hat{y} \log \hat{y}$. In both instances, our program estimated $\hat{\beta}_3 = 0.0$ so we conclude that a transformation in y is not necessary.

Example 6.8 (Drink delivery data) As noted previously, although the linear model (5.99) appears to fit the data quite well, there are a number of problems that are unresolved. First there are outliers, observations 9 and 22 and the distribution of the residuals, as shown in Figure 6.8, have a nonnormal appearance (in fact they look roughly exponential). Hence, we consider the possibility of a transformation on y to see if these discrepancies can be more readily accounted for.

For this we used the Andrews test by fitting x_1 , x_2 and $x_3 = \hat{y} \log \hat{y}$, where \hat{y} are the fitted values for the least squares fit to y given in Table 6.2. In this case $\hat{\beta}_3 = 0.3939$ and $t_3 = 3.43$ which is significant at $< 1\%$ level. Hence, we conclude that the data should be transformed before being fit by least squares. From (6.104) we used the estimate

$$\hat{\lambda} = 1 - 0.3939 = 0.6061.$$

The model

$$\frac{Y^{(\lambda)} - 1}{\hat{\lambda}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

was used to fit to the delivery times and the ANOVA table and t statistics are shown in Tables 6.5 and 6.6.

From Table 6.5 we see that F is highly significant and both β_1 and β_2 are significantly different from zero. The fit is somewhat improved over the untransformed model.

Table 6.5 ANOVA table for transformed delivery data

Source	df	Sum of Squares	Mean Squares	F	p -value
Regression	2	353.38	176.69	323.08	0.000
Residual	22	12.03	0.55		
Total	24	365.42			
		$R^2 = 0.967$	$\bar{R}^2 = 0.964$		

Table 6.6 *t* statistics for transformed delivery model

Predictor	Coefficient	S.E. Coeff.	<i>t</i> -statistic	<i>p</i> -value
constant	3.7214	0.2488	14.96	0.000
<i>x</i> ₁	0.40422	0.03874	10.44	0.000
<i>x</i> ₂	0.0037095	0.0008198	4.53	0.000

From Table 6.6 the regression equation is

$$\frac{\hat{y}^{(\hat{\lambda})} - 1}{\hat{\lambda}} = 3.72 + 0.404\beta_1 + 0.00371\beta_2.$$

To examine model assumptions we investigate the residual plots in Figure 6.14. There appears to be distinct improvements over the untransformed data. The histogram is more symmetric and both observations 9 and 22 are no longer outliers. Overall, the transformed model appears to give a better representation of the data than our original assumptions in Example 5.14.

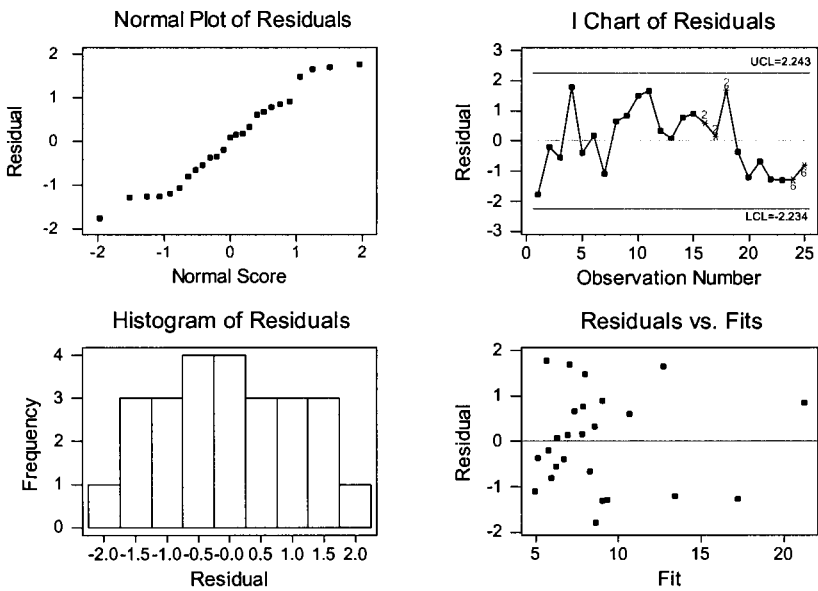


Figure 6.14: Plots of residuals for transformed drink delivery data

Example 6.9 (Tree data) Like the Longley and Hald data, the data shown in Table 6.7 have achieved a certain notoriety in the statistics literature [22, 87, 27]. The data are measurements taken on a sample of 31 black cherry trees in the Allegheny National Forrest, Pennsylvania, USA. The values of *D* are the diameters of the trees taken at a height of 4.5 feet above ground level and *H* is the height of the measured trees. *V* is the volume of the trees in cubic feet. The data were collected to provide a way for estimating the amount of timber a tree yields using its height and diameter.

Table 6.7 Allegheny National Forrest Tree data

No.	Diameter	Height	Volume	No.	Diameter	Height	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

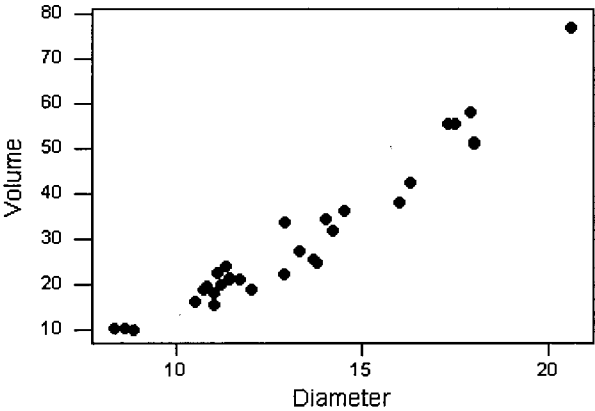


Figure 6.15: Scatter plot of volume (V) versus diameter (D)

To model the data we made scatter plots of V against H and V against D . These are shown in Figures 6.12-6.13. Linear trends are seen in both cases so we begin our analysis by fitting the linear model

$$Y = \beta_0 + \beta_1 D + \beta_2 H + \varepsilon \tag{6.105}$$

($D = x_1$, $H = x_2$) to the data and the ANOVA and t statistics are shown in Tables 6.8 and 6.9.

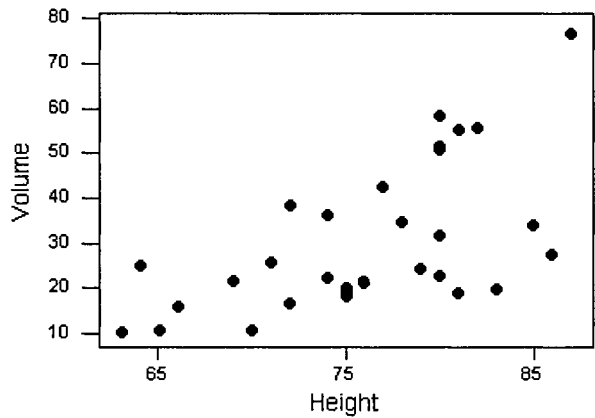


Figure 6.16: Scatter plot of volume (V) versus height (H)

Table 6.8 ANOVA table for tree data

Source	df	Sum of Squares	Mean Squares	F	p -value
Regression	2	7684.2	3842.1	254.97	0.000
Residual	28	421.9	15.1		
Total	30	8106.1			
		$R^2 = 0.948$	$\bar{R}^2 = 0.944$		

Table 6.9 t statistics for tree model

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value	VIF
constant	-57.988	8.6380	-6.71	0.000	
x_1	4.7082	0.2643	17.82	0.000	1.4
x_2	0.3393	0.1302	2.61	0.014	1.4

From these tables we see that the overall fit is highly significant ($P(F > 254.97) < 10^{-3}$) and the t values show that β_1 and β_2 are significantly different from zero.

To further examine the validity of the model we computed the residuals and various influence statistics which are shown in Table 6.10. Residual plots for \hat{t}_i are given in Figure 6.17. Two features are apparent. First, observation 31 appears as an outlier and as measured by Cook's D is the most influential. Since it is the “largest” tree, this is not surprising since observations at extreme points of the data often are the most influential. Second, the residual plot, Figure 6.18, shows a rather distinct “bowl” shaped character suggesting that a transformation might be useful.

Table 6.10 Residual \hat{t}_i , Leverage h_{ii} , Cook's D_i , and DFFITS for Tree Data

No.	\hat{t}_i	h_{ii}	D_i	DFF	No.	\hat{t}_i	h_{ii}	D_i	DFF
1	1.532	0.1158	0.0978	0.555	17	0.606	0.1313	0.0189	0.235
2	1.652	0.1472	0.1479	0.686	18	-1.860	0.1435	0.1775	-0.761
3	1.568	0.1769	0.1673	0.727	19	-1.324	0.0667	0.0407	-0.354
4	0.137	0.0592	0.0004	0.034	20	-1.106	0.2112	0.1083	-0.572
5	-0.289	0.1207	0.0040	-0.107	21	0.029	0.0358	0.0000	0.006
6	-0.368	0.1558	0.0084	-0.156	22	-1.142	0.0454	0.0205	-0.249
7	-0.159	0.1148	0.0011	-0.057	23	0.238	0.0500	0.0010	0.054
8	-0.272	0.0515	0.0014	-0.063	24	-0.946	0.1114	0.0376	-0.335
9	0.318	0.0920	0.0035	0.100	25	-0.601	0.0693	0.0092	-0.164
10	-0.075	0.0480	0.0001	-0.017	26	1.213	0.0884	0.0468	0.3778
11	0.578	0.0738	0.0091	0.163	27	0.940	0.0960	0.0314	0.306
12	-0.122	0.0481	0.0003	-0.027	28	1.347	0.1064	0.0700	0.465
13	-0.018	0.0481	0.0000	-0.004	29	-0.648	0.1098	0.0177	-0.228
14	0.209	0.0728	0.0012	0.058	30	-0.786	0.1098	0.0258	-0.276
15	-1.290	0.0377	0.0212	-0.255	31	2.766	0.2271	0.6052	1.499
16	-1.517	0.0357	0.0271	-0.292					

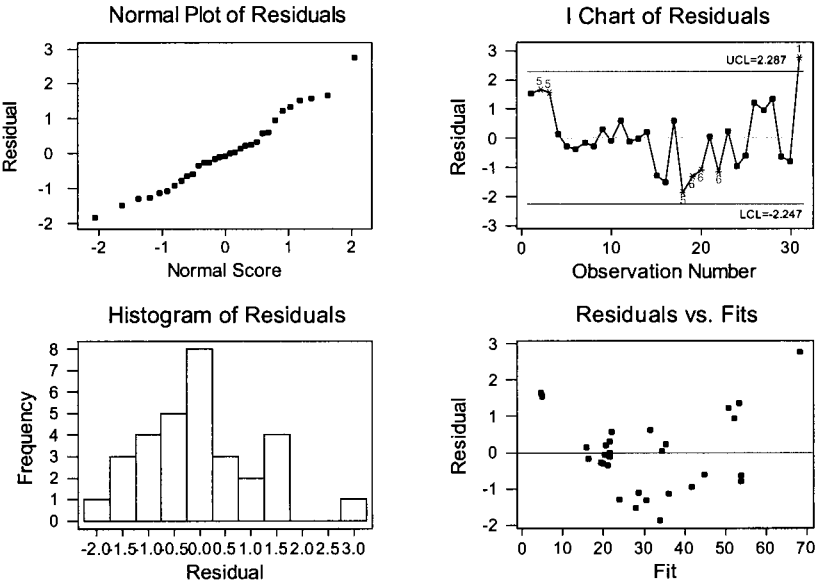


Figure 6.17: Residual plots for tree data

To substantiate this we again used Andrew’s test and this gave the value

$$\hat{\beta}_3 = 0.52953, t_3 = 5.9 \text{ and } p = 0.002.$$

Hence, β_3 is significantly different from zero. Using it,

$$\hat{\lambda} = 0.47047. \quad (6.106)$$

Fitting of the transformed model with the data transformed as $V^{(\hat{\lambda})}$ is left as an exercise. Before completing our analysis we discuss some similar results given [20] using the Box-Cox method to estimate λ . For the model

$$Y^{(\lambda)} = \beta_0 + \beta_1 D + \beta_2 H + \varepsilon \quad (6.107)$$

it was found in [20] that the Box-Cox estimate $\hat{\lambda}$ of λ is 0.307, similar to (6.106). On the basis of dimensional considerations they transformed the data using $\lambda = 1/3$ and compared the models

$$Y^{(\lambda)} = \beta_0 + \beta_{12} D^2 + \beta_2 H + \varepsilon \quad (6.108)$$

and

$$Y^{(\lambda)} = \beta_0 + \beta_{11} D + \beta_{12} D^2 + \beta_2 H + \varepsilon. \quad (6.109)$$

Generally these models improved the fit over the untransformed model (6.105) but observation 31 was still an outlier (see Exercise 6.6).

We now consider another approach to modeling the tree data which was suggested in [20] but apparently not carried out. Roughly speaking, a tree (exclusive of branches and leaves) has the appearance of a truncated cone. For a cone, its volume V is given by

$$V = \pi D^2 H / 3 \quad (6.110)$$

Hence, for a tree it seems reasonable to assume that

$$V \simeq c D^{\beta_1} H^{\beta_2} \quad (6.111)$$

where (c, β_1, β_2) are estimated from the data. To do this we take the logarithm of (6.111) giving the model ($\beta_0 = \log c$)

$$\log V = \beta_0 + \beta_1 \log D + \beta_2 \log H + \varepsilon. \quad (6.112)$$

This model was fit using least squares and the results are shown in Tables 6.11 and 6.12.

Table 6.11 ANOVA table for log transformed tree data

Source	df	Sum of Squares	Mean Squares	F	p -value
Regression	2	8.1232	4.0616	613.19	0.000
Residual	28	0.1855	0.0066		
Total	30	8.3087			
		$R^2 = 0.978$	$\bar{R}^2 = 0.976$		

Table 6.12 t statistics for log transformed tree model

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value	VIF
constant	-6.6316	0.7998	-8.29	0.000	
x_1	1.9827	0.0750	26.43	0.000	1.4
x_2	1.1171	0.2044	5.46	0.000	1.4

From these tables we see that the overall fit is highly significant ($P\{F > 613.9\} < 10^{-3}$) and again β_1 and β_2 are significantly different from zero. Notice also that $\hat{\beta}_1 =$

$1.98265 \simeq 2$ and $\hat{\beta}_2 = 1.1171$ which are close to the theoretical values $\beta_1 = 2$ and $\beta_2 = 1$ if (6.112) were exactly true. So the model not only fits the data well, but does so in a way that corresponds to the “physics” of the situation. To further investigate the validity of the model we calculated the residuals and various influence diagnostics which are given in Table 6.13.

Finally, we give residual plots for \hat{t}_i in Figure 6.18. Notice that the histogram appears quite normal and the curvature in the residual plot, Figure 6.17 is no longer present. Moreover, the I Chart shows that observation 31 is no longer an outlier but observations 15 and 18 appear to be, but are not influential. Overall, the model based on (6.112) seems to be a good representation of the data and should be useful for prediction.

Table 6.13 Residual \hat{t}_i and Influential Diagnostics for Log-transformed Tree Data

No.	\hat{t}_i	h_{ii}	D_i	DFB	No.	\hat{t}_i	h_{ii}	D_i	DFB
1	0.287	0.1514	0.0051	0.121	17	1.598	0.1164	0.1062	0.580
2	0.455	0.1672	0.0143	0.204	18	-2.326	0.1255	0.2236	-0.881
3	0.187	0.1975	0.0030	0.093	19	-0.931	0.0711	0.0222	-0.258
4	-0.132	0.0589	0.0004	-0.033	20	-0.046	0.2428	0.0002	-0.026
5	-0.557	0.1214	0.0147	-0.207	21	0.915	0.0375	0.0109	0.181
6	-0.553	0.1521	0.0187	-0.234	22	-0.848	0.0461	0.0117	-0.187
7	-0.723	0.1194	0.0240	-0.266	23	1.461	0.0539	0.0389	0.349
8	-0.552	0.0506	0.0056	-0.128	24	0.031	0.1145	0.0000	0.011
9	1.061	0.0907	0.0373	0.335	25	-0.038	0.0709	0.0000	-0.010
10	0.114	0.0465	0.0002	0.025	26	1.098	0.0855	0.0372	0.335
11	1.703	0.0722	0.0704	0.475	27	0.690	0.0916	0.0163	0.219
12	0.163	0.0464	0.0005	0.036	28	1.069	0.0982	0.0413	0.353
13	0.397	0.0464	0.0026	0.088	29	-0.676	0.1006	0.0174	-0.226
14	1.072	0.0728	0.0299	0.300	30	-0.804	0.1006	0.0244	-0.269
15	-2.258	0.0356	0.0547	-0.434	31	-0.155	0.1803	0.0018	-0.073
16	-1.920	0.0356	0.0414	-0.369					

6.5.7 Variance Equalizing Transformations

When model checking suggests that the error variances are not constant, then one should consider transforming the data, to equalize (at least approximately) the variances of ε_i , before using least squares. By definition, the Box-Cox and related transformations do this by assuming that the transformed data is $N(0, \sigma^2)$. If such a transformation is not warranted, in particular, if the data is discrete, then other approaches are necessary. We consider two approaches:

- (i) weighted least squares;
- (ii) variance stabilizing transformations.

Weighted Least Squares

Here we assume that the errors in (5.1) are normal, but $Var(\varepsilon_i) = \sigma_i^2$ depends on the i -th observation. In particular, we assume that

$$\sigma_i^2 = \sigma^2/w_i, \tag{6.113}$$

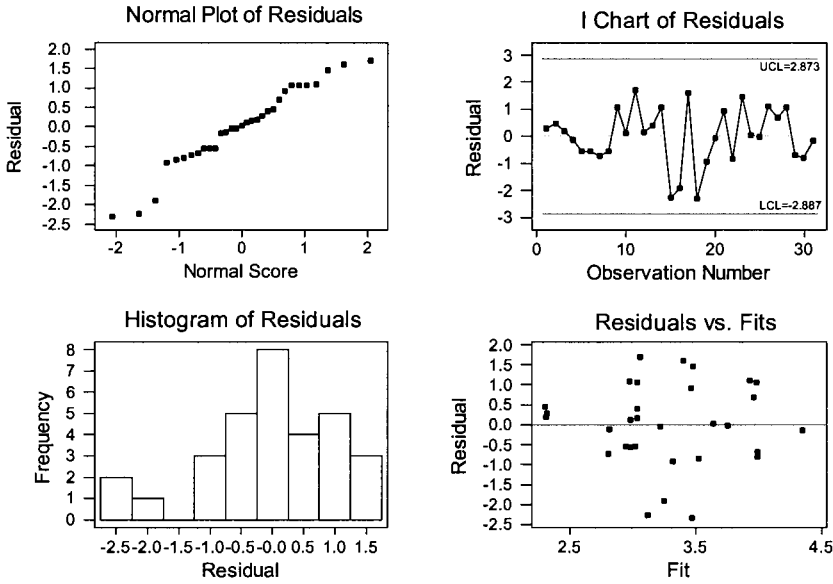


Figure 6.18: Residual plots for log-transformed model of tree data

where $w_i > 0$, $1 \leq i \leq n$, are referred to as *weights*. Our assumption then is that the model is of the form (5.1) except that $\varepsilon_i \sim N(0, \sigma^2/w_i)$, $1 \leq i \leq n$. If the weights are known, then we can estimate β_j , $0 \leq j \leq m$, and σ^2 by MLE as follows.

Letting

$$\mu_i = \sum_{j=0}^m \beta_j x_{ij} \quad (6.114)$$

then the likelihood function of Y_i , $1 \leq i \leq n$, is given by

$$L = \frac{1}{(2\pi\sigma)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=0}^n w_i (y_i - \mu_i)^2 \right] \quad (6.115)$$

and minimizing L with respect to β and σ^2 enables us to obtain $\hat{\beta}$ and $\hat{\sigma}^2$.

To simplify the arithmetic, observe that

$$w_i (y_i - \mu_i)^2 = (\sqrt{w_i} y_i - \sqrt{w_i} \mu_i)^2 \quad (6.116)$$

which gives

$$w_i (y_i - \mu_i)^2 = (z_i - r_i)^2 \quad (6.117)$$

where $z_i = \sqrt{w_i} y_i$, $1 \leq i \leq n$, and

$$r_i = \sum_{j=0}^m \beta_j \sqrt{w_i} x_{ij} = \sum_{j=0}^m \beta_j v_{ij} \quad (6.118)$$

where $v_{ij} = \sqrt{w_i}x_{ij}$, $1 \leq i \leq n$, $0 \leq j \leq m$, then

$$L = \frac{1}{(2\pi\sigma)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=0}^n \langle z_i - r_i \rangle^2 \right]. \quad (6.119)$$

Letting

$$\begin{aligned} \mathbf{V} &= \begin{bmatrix} \sqrt{w_1} & \sqrt{w_1}x_{11} & \cdots & \sqrt{w_1}x_{1m} \\ \sqrt{w_2} & \sqrt{w_2}x_{21} & \cdots & \sqrt{w_2}x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_n} & \sqrt{w_n}x_{n1} & \cdots & \sqrt{w_n}x_{nm} \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_n} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \\ &= \sqrt{\mathbf{W}}\mathbf{X} \end{aligned} \quad (6.120)$$

where $\sqrt{\mathbf{W}} = \text{diag}(\sqrt{w_1}, \sqrt{w_2}, \dots, \sqrt{w_n})$, then L is the likelihood function for the model

$$\mathbf{Z} = \mathbf{V}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6.121)$$

where $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{z} = (\sqrt{w_1}y_1, \sqrt{w_2}y_2, \dots, \sqrt{w_n}y_n)^T$. Thus, $\boldsymbol{\beta}$ can be estimated by least squares as before with the proviso that the model (6.121) does not have an intercept since the zero-th column of \mathbf{V} is $(\sqrt{w_1}, \sqrt{w_2}, \dots, \sqrt{w_n})^T$. Since the least squares estimates of $\boldsymbol{\beta}$ are not affected by this, it follows that $\boldsymbol{\beta}$ is estimated by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_w &= (\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \mathbf{z} \\ &= \left[(\sqrt{\mathbf{W}}\mathbf{X})^T \sqrt{\mathbf{W}}\mathbf{X} \right]^{-1} (\sqrt{\mathbf{W}}\mathbf{X})^T \sqrt{\mathbf{W}}\mathbf{y} \\ &= (\mathbf{X}^T \sqrt{\mathbf{W}} \sqrt{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \sqrt{\mathbf{W}} \sqrt{\mathbf{W}} \mathbf{y} \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \end{aligned} \quad (6.122)$$

Also, the MLE $\hat{\sigma}_w^2$ of σ^2 is given by

$$\hat{\sigma}_w^2 = \frac{1}{n} \sum_{i=1}^n w_i (y_i - \hat{\mu}_i)^2 \equiv \frac{SSE_w}{n} \quad (6.123)$$

where $\hat{\mu}_i = \langle \hat{\boldsymbol{\beta}}, \mathbf{x}_i \rangle$, and SSE_w is the *weighted sum of squares*.

From (6.122)

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_w) &= E \left[(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \right] \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{W} \mathbf{X}) \boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned} \quad (6.124)$$

so that $\hat{\beta}_w$ is an unbiased estimator of β . Also, since SSE_w is the residual sum of squares from fitting (6.121) it follows from Theorem 5.1 that

$$E \left(\frac{SSE_w}{n - m - 1} \right) = \sigma^2 \quad (6.125)$$

so that

$$s_w^2 = \frac{SSE_w}{n - m - 1} \quad (6.126)$$

is an unbiased estimator of σ^2 . Then $\sqrt{s_w^2}$ is taken as the estimate of σ . Note that (6.122) is not the OLS estimator $\hat{\beta}$ of β in the model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. From the Gauss-Markov theorem $\hat{\beta}_w$ is the BLUE estimator of β in (6.121), hence the OLS estimator of β is *less efficient* than $\hat{\beta}_w$. The estimate of β in (6.121) is usually called the *weighted least squares estimator* (WLS) and is generally preferred to $\hat{\beta}_{OLS}$ when $\varepsilon_i \sim N(0, \sigma^2/w_i)$, $1 \leq i \leq n$.

A number of additional properties of $\hat{\beta}_w$ are given next.

Theorem 6.4 *Let $\hat{\beta}_w$ be the weighted least squares estimator of β when $\Sigma(\varepsilon) = \sigma^2 \text{diag}(1/w_1, 1/w_2, \dots, 1/w_n)$ in (6.122). Then,*

$$(i) \Sigma(\hat{\beta}_w) = \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1};$$

(ii) *Let δ_i be the i -th diagonal element of $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$. If $\varepsilon_i \sim N(0, \sigma^2/w_i)$, $1 \leq i \leq n$, then $(\hat{\beta}_{w,i} - \beta_i) / \sigma \sqrt{\delta_i}$ is $N(0, 1)$ and*

$$T_i = \frac{\hat{\beta}_{w,i} - \beta_i}{s_w \sqrt{\delta_i}} \quad (6.127)$$

has a t -distribution with $n - m - 1$ degrees of freedom.

(iii) *Let $\hat{\mathbf{Y}}_w = \mathbf{X}\hat{\beta}_w$, then $E(\hat{\mathbf{Y}}_w) = \mathbf{X}\beta$, so that $\hat{\mathbf{Y}}$ is an unbiased estimator of $\mathbf{X}\beta$ and*

$$\Sigma(\hat{\mathbf{Y}}_w) = \sigma^2 \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \quad (6.128)$$

(iv) *Let $\hat{\varepsilon}_w = \mathbf{Y} - \hat{\mathbf{Y}}_w$ be the residuals from $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ and $\hat{\varepsilon}_w^* = \mathbf{Z} - \hat{\mathbf{Z}} = \mathbf{Z} - \mathbf{V}\hat{\beta}_w$ be the residuals from the transformed model (6.121). Then,*

$$\hat{\varepsilon}_w^* = \sqrt{\mathbf{W}} \hat{\varepsilon}_w \quad (6.129)$$

and $E(\varepsilon_w^) = E(\varepsilon_w) = 0$.*

(v) *Let $\mathbf{H}_w = \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$ be the weighted hat matrix. Then*

$$\hat{\varepsilon}_w = (\mathbf{I} - \mathbf{H}_w) \varepsilon \quad (6.130)$$

and

$$\Sigma(\hat{\varepsilon}_w) = \sigma^2 (\mathbf{I} - \mathbf{H}_w) \mathbf{W}^{-1}. \quad (6.131)$$

Hence,

$$\Sigma(\hat{\varepsilon}_w^*) = \sigma^2 \mathbf{W}^{1/2} (\mathbf{I} - \mathbf{H}_w) \mathbf{W}^{-1/2}. \quad (6.132)$$

Proof. (i) Since $\hat{\beta}_w = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$, it follows from Theorem 4.16 that

$$\begin{aligned} \Sigma(\hat{\beta}_w) &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \Sigma(\mathbf{Y}) \left[(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \right]^T \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \sigma^2 \mathbf{W}^{-1} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}. \end{aligned} \quad (6.133)$$

(ii) This follows because $\text{Var}(\hat{\beta}_{w,i}) = \sigma^2 \delta_i$ and from Theorem 5.4 s_w^2 is independent of $\hat{\beta}_{w,i}$ so that this follows as in Theorem 5.4.

(iii) Since $\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta}_w$ $E(\hat{\mathbf{Y}}) = \mathbf{X} E(\hat{\beta}_w) = \mathbf{X} \beta$. Also

$$\Sigma(\hat{\mathbf{Y}}_w) = \mathbf{X} \Sigma(\hat{\beta}_w) \mathbf{X}^T = \sigma^2 \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T. \quad (6.134)$$

(iv) From (6.121) $\mathbf{Z} = \sqrt{\mathbf{W}} \mathbf{Y}$ and $\mathbf{V} = \sqrt{\mathbf{W}} \mathbf{X}$, so that

$$\begin{aligned} \hat{\varepsilon}_w^* &= \mathbf{Z} - \hat{\mathbf{Z}} = \sqrt{\mathbf{W}} \mathbf{Y} - \sqrt{\mathbf{W}} \mathbf{X} \hat{\beta}_w \\ &= \sqrt{\mathbf{W}} (\mathbf{Y} - \mathbf{X} \hat{\beta}_w) = \sqrt{\mathbf{W}} \hat{\varepsilon}. \end{aligned} \quad (6.135)$$

Since $E(\hat{\varepsilon}) = E(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{X} \beta - \mathbf{X} \beta = \mathbf{0}$, it follows that $E(\hat{\varepsilon}_w^*) = \sqrt{\mathbf{W}} E(\hat{\varepsilon}) = \mathbf{0}$.

(v) By definition, $\hat{\mathbf{Y}}_w = \mathbf{X} \hat{\beta}_w = \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$ and $\mathbf{Y} = \mathbf{X} \beta + \varepsilon$ so that

$$\begin{aligned} \hat{\varepsilon}_w &= \mathbf{Y} - \hat{\mathbf{Y}}_w = \mathbf{X} \beta + \varepsilon - \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} (\mathbf{X} \beta + \varepsilon) \\ &= \mathbf{X} \beta - \mathbf{X} \beta + \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \varepsilon + \varepsilon \\ &= \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \right] \varepsilon \\ &= (\mathbf{I} - \mathbf{H}_w) \varepsilon. \end{aligned} \quad (6.136)$$

Using the fact that $\mathbf{H}_w^2 = \mathbf{H}_w$ (it is a projection matrix) (6.131) and (6.132) follow by some tedious algebra. We leave the details to the reader. ■

From Theorem 6.4 and the prior results it follows that the WLS estimator $\hat{\beta}_w$ of β is obtained and its standard error can be obtained by using the transformed model (6.121). Moreover, significance tests for β can be done using the transformed model as well. However, one must be careful in performing F tests and computing R^2 .

Since the transformed model does not have an intercept, the usual decomposition of the sum of squares does not hold. Hence the F statistic for the transformed model cannot be defined in the usual way and the issue of R^2 is analogous to that considered for SLR through the origin. However, since the “extra sum of squares” principle is valid whether or not the model has an intercept, an F test for the overall significance of the regression can be obtained from the statistic

$$F_w = \frac{(SSE_R - SSE_F)/m}{s_w^2} \quad (6.137)$$

where SSE_F is the error sum of squares s_w^2 from the full model and SSE_R is the residual sum of squares from the reduced transformed model

$$\mathbf{Z} = \mathbf{V}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \quad (6.138)$$

where $\mathbf{V}_0 = (\sqrt{w_1}, \sqrt{w_2}, \dots, \sqrt{w_n})^T$. If the errors are normal, under the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0 \quad (6.139)$$

and F_w has an F distribution with $(m, n - m - 1)$ degrees of freedom, so we reject H_0 at level α if

$$F_w > f_{\alpha, m, n-m-1}. \quad (6.140)$$

This issue of an appropriate goodness of fit measure is somewhat controversial, but a natural choice, as for SLR, is to define

$$R^2 = \rho^2 \quad (6.141)$$

where ρ is the sample correlation coefficient of $\hat{\mathbf{z}}$ and \mathbf{z} . This is consistent with our approach for SLR and reduces to R^2 when $\mathbf{W} = \mathbf{I}$.

For model checking, it is necessary to consider appropriate residual plots. Since we have two sets of residuals, $\hat{\varepsilon}_i$ from the original model (5.1) and $\hat{\varepsilon}_i^*$ from the transformed model, we have two choices. As for the case of constant variance, one can also consider standardized and studentized residuals obtained by dividing $\hat{\varepsilon}_i^*$ and $\hat{\varepsilon}_i$ by their standard errors. Using (vi) in Theorem 6.4 and (6.120) it can be shown that the standardized, hence, studentized residuals are the same. However, we must be careful in determining what values we should choose for these to be orthogonal to the independent variables \mathbf{x}_i , $1 \leq i \leq n$, or predicted values. This will be true if we plot ε_i^* from the transformed model against the columns of \mathbf{V} and the predicted values $\hat{\mathbf{z}}$. For example, using (6.129)

$$\sum_{i=1}^n \hat{z}_i \hat{\varepsilon}_i^* = 0. \quad (6.142)$$

Since $\varepsilon_i^* = \sqrt{w_i} \hat{\varepsilon}_i$ and $\hat{z}_i = \sqrt{w_i} \hat{y}_i$ using these in (6.142) gives

$$\sum_{i=1}^n w_i \hat{\varepsilon}_i \hat{y}_i = 0. \quad (6.143)$$

Thus the residuals $\hat{\varepsilon}_i$ from the original model are not orthogonal to the predicted values \hat{y}_i , hence plotting $\hat{\varepsilon}_i$ against \hat{y}_i will give a confusing plot. However (6.143) shows that plotting $\sqrt{w_i} \hat{\varepsilon}_i$ against $\sqrt{w_i} \hat{y}_i$ is appropriate. Similar observations hold for plots of $\hat{\varepsilon}_i$ against the variables \mathbf{x}_i , $1 \leq i \leq n$.

One can also examine the leverage of \hat{y}_i , $1 \leq i \leq n$, and influence diagnostics using the transformed model.

Weights Unknown

In general the weights in (6.114) are unknown and need to be estimated from the data along with β_j , $0 \leq j \leq m$, and σ^2 . Unfortunately in this generality this is not possible

since there are more unknowns than observations. However, in many situations one may have some further knowledge about the error distribution, which enables one to do this.

For example, suppose we have an experiment with a binary response $y = 0$ or $y = 1$. For example, suppose the effect of a prescription drug depends on the dose of the drug. As a result of giving dose “ x ” to n_x patients if we assume that the sample of patients is a random sample of size n_x , then the number of recoveries r_x is a binomial random variable Y_x with

$$P\{Y_x = r_x\} = \binom{n_x}{r_x} (p_x)^{r_x} (1 - p_x)^{n_x - r_x} \quad (6.144)$$

where p_x is the probability of a recovery at dose x . From (6.144)

$$E(Y_x) = n_x p_x \quad \text{and} \quad \text{Var}(Y_x) = n_x p_x (1 - p_x). \quad (6.145)$$

Unless n_x and p_x are constant, the variances of Y_x are unequal. If, for simplicity, we assume that Y_x depends linearly on x , (intuitively, one would expect the effect of the drug to increase as the dose increases), then

$$Y_x = \beta_0 + \beta_1 x + \varepsilon_x \quad (6.146)$$

where $\text{Var}(Y_x) = n_x p_x (1 - p_x)$.

To estimate β_0, β_1 we can use WLS and minimize

$$\sum_{i=1}^q \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{n_{x_i} p_{x_i} (1 - p_{x_i})} \quad (6.147)$$

where $y_i = r_{x_i}$ is the number of recoveries at dose $x_i, 1 \leq i \leq q$. However, p_x are unknown, so we seem to have an impasse. In this instance a reasonable estimate of p_x is r_x/n_x so using this in (6.147) we can estimate β_0, β_1 by minimizing

$$\sum_{i=1}^q \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{n_{x_i} (r_{x_i}/n_{x_i}) (1 - r_{x_i}/n_{x_i})} = \sum_{i=1}^q \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{y_i (1 - y_i/n_{x_i})} \quad (6.148)$$

which is now the SSE for a model with known variances $\sigma_i^2 = y_i (1 - y_i/n_{x_i})$.

This process can be iterated. Suppose $\hat{y}_i^{(1)}/n_{x_i}$ are the predicted proportions from the WLS fit above. Then we can estimate the variance by $\hat{y}_i^{(1)} (1 - \hat{y}_i^{(1)}/n_{x_i})$. These can then be used as new weights and β_0, β_1 reestimated. Obviously, this process can be repeated, and if it converges it allows one to estimate the parameters by WLS even if the weights are unknown. It is interesting to note that if this process converges, then the limiting values $\hat{\beta}_0, \hat{\beta}_1$ are the MLEs of β_0, β_1 . Hence, for this model, MLE is equivalent to *iteratively reweighted least squares* [14]. We shall return to this matter in Chapter 7.

Example 6.10 The data in Table 6.14 was collected for the purpose of determining if the probability of owning a home is related to income. In column 1 are the incomes (in dollars) of 20 people (x) and column 2 lists homeowner status, 1 for an owner and 0 if not (y). We consider whether there is a linear relation between y and x . Hence we propose a model of the form

$$Y_x = \beta_0 + \beta_1 x + \varepsilon_x \quad (6.149)$$

where Y_x is a Bernoulli random variable and $E(\varepsilon_x) = 0$ so that

$$E(Y_x) = \beta_0 + \beta_1 x = P\{Y_x = 1\}.$$

(6.150)

Hence $Var(Y_x) = (\beta_0 + \beta_1 x)(1 - \beta_0 - \beta_1 x)$ and Y_x does not have constant variance. Hence, the Gauss-Markov theorem cannot be used to obtain optimal linear estimates of (β_0, β_1) . Of course, if we know the weights in (6.113) then we could used weighted least squares to obtain these estimates. Since these are unknown, we proceed in an iterative fashion as indicated above. We first fit (6.149) by OLS and then we estimate the weights by

$$w_i = [\hat{y}_i(1 - \hat{y}_i)]^{-1}$$

(6.151)

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates of β_0 and β_1 , $1 \leq i \leq 20$. This was done and the fitted values and residuals are shown in the first two columns of Table 6.17.

Table 6.14 Homeowner data

Obs. No.	x	y	Obs. No.	x	y
1	8300	0	11	18700	1
2	21200	1	12	10100	0
3	9100	0	13	19500	1
4	13400	1	14	8000	0
5	17700	0	15	12000	1
6	23000	0	16	24000	1
7	11500	1	17	21700	1
8	10800	0	18	9400	0
9	15400	1	19	10900	0
10	22400	1	20	22800	1

Table 6.15 ANOVA table for Homeowner Data

Source	df	Sum of Squares	Mean Squares	F	p -value
Regression	1	1.6603	1.6603	9.08	0.007
Residual	18	3.2897	0.1828		
Total	19	4.9500			
		$R^2 = 0.3350$	$\bar{R}^2 = 0.2980$		

Table 6.16 t statistics for Homeowner data

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value
constant	-0.2501	0.2821	-0.89	0.387
x_1	0.0 ⁴ 5163	0.0 ⁴ 1713	3.01	0.007

From Table 6.15 it appears that the overall regression is significant with β_1 appearing to be significantly different from zero. However, it is not clear that these statistics can be given their usual interpretation since the errors in (6.149) are not normal. To get further confirmation of this we made residual plots of the ordinary residuals against \hat{y}_i , a histogram of residuals, a normal plot and an I Chart. Examining Figure 6.19. We see residuals that look roughly normal, but the residual plot is rather different. One can clearly see the effect of the nonconstant variance of the residuals. The residuals systematically decrease in parallel. To emphasize this, we show a plot of the absolute values of $\hat{\varepsilon}_i$ against \hat{y}_i . Such intersecting (or X-shaped) plots are characteristic of errors

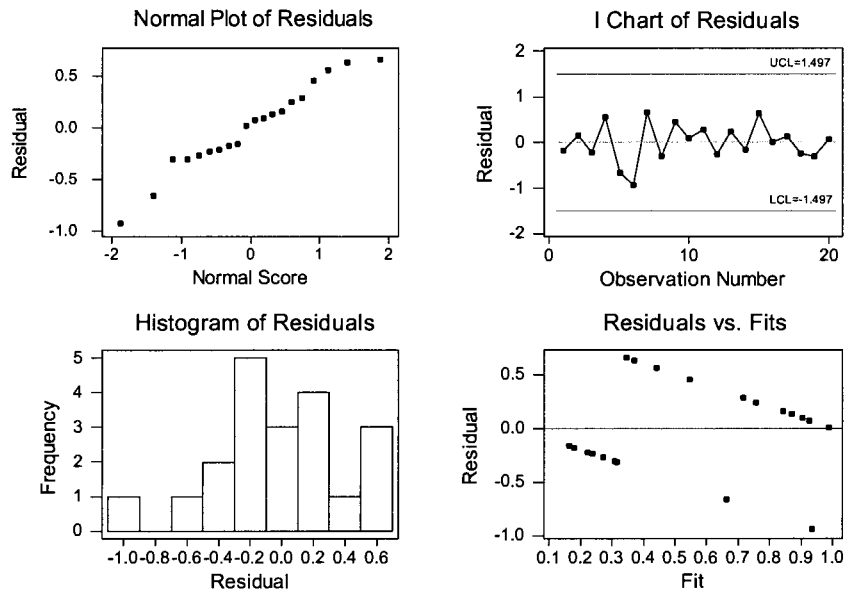


Figure 6.19: Plots of ordinary residuals for homeowners data

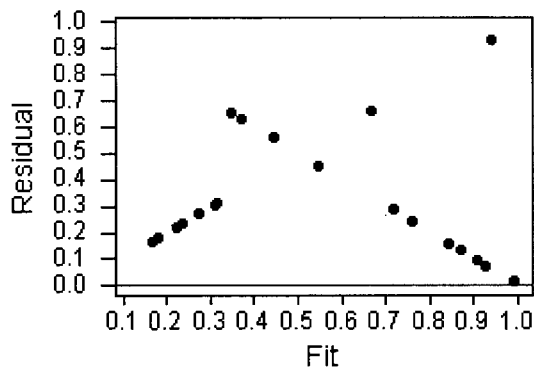


Figure 6.20: Plot of ordinary residuals $|\hat{\varepsilon}_i|$ versus \hat{y}_i

associated with binary responses. Clearly, making inferences about the estimated model using normal statistics should be treated cautiously.

Table 6.17 Residuals and Weighted Residuals

Obs. No.	OLS Fit	OLS Res	Weight	WLS Fit	WLS Res
1	0.1785	-0.1785	6.8197	0.141879	-0.141879
2	0.8446	0.1554	7.6179	0.797688	0.202312
3	0.2198	-0.2198	5.8313	0.182549	-0.182549
4	0.4418	0.5582	4.0549	0.401152	0.598848
5	0.6639	-0.6639	4.4812	0.619755	-0.619755
6	0.9375	-0.9375	17.0699	0.889196	-0.889196
7	0.3437	0.6563	4.4331	0.304560	0.695440
8	0.3076	-0.3076	4.6954	0.268973	-0.268973
9	0.5451	0.4549	4.0328	0.502828	0.497172
10	0.9065	0.0935	11.8020	0.858693	0.141307
11	0.7155	0.2845	4.9124	0.670593	0.329407
12	0.2714	-0.2714	5.0567	0.233387	-0.233387
13	0.7568	0.2432	5.4331	0.711263	0.288737
14	0.1630	-0.1630	7.3296	0.126627	-0.126627
15	0.3695	0.6305	4.2922	0.329979	0.670021
16	0.9891	0.0109	93.1473	0.940034	0.059966
17	0.8704	0.1296	8.8643	0.823107	0.176893
18	0.2353	-0.2353	5.5577	0.197800	-0.197800
19	0.3127	-0.3127	4.6526	0.274057	-0.274057
20	0.9272	0.0728	14.8121	0.879028	0.120972

To see if one gets improvement using WLS the model (6.149) was refit using the weights in column 4 of Table 6.17. A direct comparison could not be made between OLS and WLS so we computed the test statistics as if this was an OLS fit. The ANOVA table, t statistics and residual plots are shown in Tables 6.18-6.19 and Figures 6.21-6.22.

Table 6.18 ANOVA table for Homeowner data using WLS

Source	df	Sum of Squares	Mean Squares	F	p -value
Regression	1	18.946	18.946	13.25	0.002
Residual	18	25.733	1.430		
Total	19	44.680			
		$R^2 = 0.424$	$\bar{R}^2 = 0.392$		

Table 6.19 t statistics for Homeowner data using WLS

Predictor	Coefficient	S.E. Coeff.	t -statistic	p -value
constant	-0.2801	0.2878	-0.97	0.343
x_1	0.045084	0.041396	3.64	0.002

From Tables 6.18-6.19 we see that the WLS fit is quite similar to the OLS results, although the standard errors of the coefficients are a little smaller. However, the plot of $|\hat{\epsilon}_i|$ against \hat{y}_i still has a characteristic X shape. Again the nonconstant variance and non-normality of the errors makes inferences problematical. We shall return to this matter in Chapter 7.

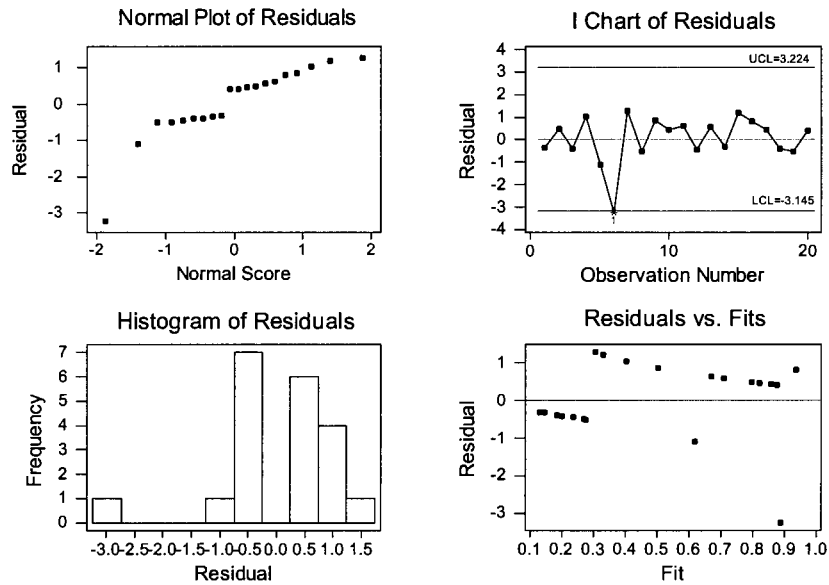


Figure 6.21: Plots of weighted residuals for homeowners data

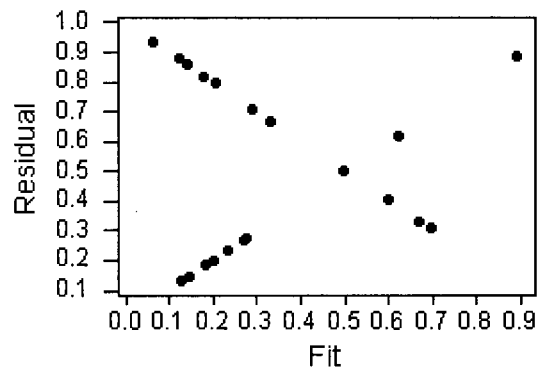


Figure 6.22: Plot of weighted $|\hat{\epsilon}_i|$ versus \hat{y}_i

6.5.8 Variance Stabilizing Transformations

Another approach to equalizing variances is to try to transform the observations y_i so that the transformed values $z_i = g(y_i)$ have approximately equal variance. As a particular case, the Box-Cox method does this, but more general transformations may be needed than power transformations.

To simplify matters, assume that the variance of Y_i is a function of $\mu_i = E(Y_i)$. For the binomial distribution

$$\text{Var}(Y_i) = n_i p_i (1 - p_i). \quad (6.152)$$

However, $E(Y_i) = n_i p_i = \mu_i$ so that $p_i = E(Y_i)/n_i$ and

$$\text{Var}(Y_i) = n_i (\mu_i/n_i) (1 - \mu_i/n_i) = (n_i - \mu_i) (\mu_i/n_i) = g(\mu_i). \quad (6.153)$$

With this assumption we will try to find g such that $\text{Var}[g(\mu_i)]$ does depend on i . Assuming, Y_i does not vary considerably from μ_i we use Taylor's series to get

$$g(Y_i) \simeq g(\mu_i) + (Y_i - \mu_i) g'(\mu_i). \quad (6.154)$$

Hence,

$$\text{Var}[g(Y_i)] \simeq [g'(\mu_i)]^2 \text{Var}(Y_i). \quad (6.155)$$

If $\text{Var}(Y_i) = f(\mu_i)$, then

$$\text{Var}[g(Y_i)] \simeq [g'(\mu_i)]^2 f(\mu_i). \quad (6.156)$$

For this to be independent of i , we must have

$$[g'(\mu_i)]^2 f(\mu_i) = \sigma^2 \quad (6.157)$$

or

$$g'(\mu_i) = \sigma / \sqrt{f(\mu_i)}. \quad (6.158)$$

Letting $\mu \equiv \mu_i$ g can be found by integrating (6.158) giving

$$g(\mu) = \sigma \int^{\mu} \frac{dx}{\sqrt{f(\mu)}}. \quad (6.159)$$

Example 6.11 If Y_i is a binomial random variable, then $f(\mu) = \mu(n - \mu)/n$, and g is given by

$$g(\mu) = \int^{\mu} \frac{dx}{\sqrt{\mu(n - \mu)/n}} = \int^{\mu} \frac{\sqrt{n} dx}{\sqrt{\mu(n - \mu)}}. \quad (6.160)$$

Letting $ny = x$, (6.160) becomes

$$\begin{aligned} \int^{\mu/n} \frac{ndy\sqrt{n}}{\sqrt{ny(n - ny)}} &= \int^{\mu/n} \frac{\sqrt{n} dy}{\sqrt{y(1 - y)}} \\ &= \sqrt{n} \int^{\mu/n} \frac{dy}{\sqrt{y(1 - y)}} \\ &= 2\sqrt{n} \sin^{-1}(\sqrt{y/n}). \end{aligned} \quad (6.161)$$

For proportion data this gives

$$Z_i = g(Y_i) = 2\sqrt{n_i} \sin^{-1} \left(\sqrt{Y_i} / n_i \right) \quad (6.162)$$

as an appropriate transformation to equalize variance. Transforming the data this way indicates that an ordinary least squares estimation procedure can be used to estimate β_0, β_1 .

Example 6.12 As we assumed previously it was suggested in OzDASL (see Ex. 6.1) that a gamma error distribution might be a better choice than a normal error model for the drink delivery data. Also, our discussion of the Box-Cox method suggested that a transformation is appropriate. Here we examine the possibility of using a variance stabilizing transformation. Since an exponential random variable is the simplest gamma random variable, we examine this possibility.

Hence, we assume that Y_i has a density

$$f(y_i) = \lambda \exp(-\lambda_i y_i). \quad (6.163)$$

In this case

$$E(Y_i) = \int_0^\infty \lambda_i y_i e^{-\lambda_i y_i} dy_i. \quad (6.164)$$

Letting $z_i = \lambda_i y_i$

$$E(Y_i) = \int_0^\infty (z_i / \lambda_i) e^{-z_i} dz_i = \Gamma(2) / \lambda_i = 1 / \lambda_i = \mu_i. \quad (6.165)$$

Similarly,

$$E(Y_i^2) = 2 / \lambda_i^2. \quad (6.166)$$

Hence,

$$\text{Var}(Y_i) = E(Y_i^2) - [E(Y_i)]^2 = 1 / \lambda_i^2 = \mu_i^2. \quad (6.167)$$

Using this in (6.156) we choose μ_i by

$$[g'(\mu_i)]^2 / \mu_i^2 = \sigma^2 \quad (6.168)$$

or

$$g'(\mu_i) / \mu_i = \sigma. \quad (6.169)$$

Integrating gives

$$g' = \int^{\mu_i} (\sigma^2 / \mu_i) d\mu_i = \sigma^2 \log \mu_i. \quad (6.170)$$

Hence, this suggests transforming the drink delivery data by

$$\log y = \beta_0 + \beta_1 x_1 + \varepsilon. \quad (6.171)$$

One can apply this model to fit the data and compare the outputs with the results from the untransformed data. We leave the details to the reader.

6.6 Correlated Errors

As we have observed in a number of places residual plots can sometimes be used to detect correlation between the errors in the GLM. Typically these will show up as systematic oscillations in plots of residuals against fitted values. Most commonly, this behavior occurs for data taken over time, such as population or economic data.

If the correlation coefficient ρ is positive, then the oscillations are slow as shown for the Clark County population data and the Longley data. When the correlation is negative, then the oscillations tend to have much short on periods since successive residuals tend to have opposite signs. When the residuals are correlated, the independence assumptions of the GLM fail so it is important to identify its occurrence and remedy it if possible. This topic is a separate branch of statistics usually called *time series analysis* and is generally beyond the scope of this text. In recent years models with spatial correlation have become increasingly important in many applied areas. This topic, usually called *kriging*, has many features in common with regression analysis - but again details are beyond the scope of the text.

If autocorrelation is present and the errors have constant variance then it follows that:

- (i) the OLS estimator of β in (5.11) is unbiased, but the Gauss-Markov theorem does not hold so the least squares estimates no longer have minimum variance;
- (ii) $MSE = SSE/(n - m - 1)$, the estimate of σ^2 , may be substantially smaller than the true value of σ^2 and give a false impression of accuracy;
- (iii) as a consequence of (ii) t values may be inflated and inferences and confidence intervals for the parameters may give a false impression of precision;
- (iv) since the errors are dependent, F and t tests are not strictly valid even if the errors are normal.

6.6.1 The Durbin-Watson Statistic

Detecting and correcting for autocorrelation is a rather complex subject so we will limit our discussion to the case where the observations are taken in time $t = 1, 2, \dots, n$ and the errors $\varepsilon_t, 1 \leq t \leq n$, satisfy the *first order autoregressive* condition

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad |\rho| < 1 \quad (6.172)$$

where ρ is the *autocorrelation coefficient* and $u_t, t \geq 1$, are independent normal random variables with constant variance σ^2 and u_t is independent of $\varepsilon_t, t \geq 1$. Generally, for time series data $\rho > 0$ (positive autocorrelation). To test the hypothesis $H_0 : \rho = 0$ against $H_1 : \rho > 0$ one usually uses the *Durbin-Watson statistic*

$$d = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2} \quad (6.173)$$

where $\hat{\varepsilon}_t, 1 \leq t \leq n$, are the residuals estimated from the least squares fit to (5.1). If one concludes that $\rho > 0$, then ρ is estimated by

$$\hat{\rho} = \frac{\sum_{t=2}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^n \hat{\varepsilon}_t^2}. \quad (6.174)$$

From (6.173) and (6.174) one can show that $d \simeq 2(1 - \hat{\rho})$. In fact, the numerator in d is

$$\begin{aligned} \sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2 &= \sum_{t=2}^n \hat{\varepsilon}_t^2 + \sum_{t=2}^n \hat{\varepsilon}_{t-1}^2 - 2 \sum_{t=2}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-1} \\ &\simeq 2 \sum_{t=2}^n \hat{\varepsilon}_t^2 - 2 \sum_{t=2}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-1} \\ &= 2 \left(\sum_{t=2}^n \hat{\varepsilon}_t^2 - \sum_{t=2}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-1} \right). \end{aligned} \quad (6.175)$$

Again, using the approximation $\sum_{t=2}^n \hat{\varepsilon}_t^2 \simeq \sum_{t=2}^n \hat{\varepsilon}_{t-1}^2$ in (6.175) it follows from (6.174) and (6.175) that

$$d \simeq 2(1 - \hat{\rho}). \quad (6.176)$$

Using the Cauchy-Schwarz inequality (see Exercise 6.8) $\sum_{t=2}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}$ lies approximately between ± 1 so that d lies between ± 4 . Hence, values of $\hat{\rho}$ near one gives d near zero while if $\hat{\rho} \simeq 0$, $d \simeq 2$. Hence, small values of d lead us to accept $H_1 : \rho > 0$ while $d \simeq 2$ leads us to accept $H_0 : \rho = 0$. The critical values for the test were given by Durbin and Watson in [29, 30, 31] and some values are reproduced in Table A.5.

The formal test procedure is given as follows:

- (i) calculate d in (6.173) using residuals from the least squares fit of (5.1)
- (ii) obtain critical values (d_L, d_U) for an appropriate sample size n and number of variables $m + 1$ in Table A.5. Then,
 - (a) reject H_0 if $d < d_L$
 - (b) do not reject H_0 if $d > d_U$
 - (c) if $d_L < d < d_U$, the test is inconclusive.

To test for negative ρ , apply the Durbin-Watson test to $d' = 4 - d$.

6.6.2 Correcting for Autocorrelation

If one concludes that the errors are autocorrelated then there are a number of procedures which can be used to correct for this possibility. Here we discuss two methods, the *Cochrane-Orcutt* method [17, 87] and one due to Hildreth and Lu [72].

In the Cochrane-Orcutt procedure the model is transformed using (6.175) to produce a model with uncorrelated errors. For simplicity, we illustrate the approach for the simple regression model ($t = \text{"time"}$)

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (6.177)$$

where ε_t satisfies (6.172).

Now

$$\begin{aligned} Y_t - \rho Y_{t-1} &= \beta_0 + \beta_1 x_t + \varepsilon_t - \rho(\beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}) \\ &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + \varepsilon_t - \rho \varepsilon_{t-1} \\ &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + u_t \end{aligned} \quad (6.178)$$

where by assumption, u_t are independent $N(0, \sigma^2)$ random variables.

If we let $x_t^* = x_t - \rho x_{t-1}$, $Y_t^* = Y_t - \rho Y_{t-1}$, $\beta_0^* = \beta_0(1 - \rho)$, $\beta_1^* = \beta_1$, then (6.178) takes the form

$$Y_t^* = \beta_0^* + \beta_1^* x_t^* + u_t, \quad 2 \leq t \leq n. \quad (6.179)$$

Now (6.179) is in the form of the GLM and the parameters (β_0^*, β_1^*) can be estimated by least squares. However, since we need to know ρ to do this, generally one has to proceed in an iterative fashion to obtain the required estimates. A procedure for doing this follows:

- (i) fit (6.177) using OLS and obtain the residuals $\hat{\varepsilon}_t, 1 \leq t \leq n$;
- (ii) estimate ρ from (6.174) or if your program produces the Durbin-Watson statistic as output, then using (6.174) a quick estimate of ρ is given by

$$\hat{\rho} = 1 - d/2; \quad (6.180)$$

- (iii) construct the variables $y_t - \hat{\rho}y_{t-1}$ and $x_t - \hat{\rho}x_{t-1}$;
- (iv) regress $y_t - \hat{\rho}y_{t-1}$ on $x_t - \hat{\rho}x_{t-1}$ to estimate β_0^*, β_1^* . Then let $\hat{\beta}_1 = \hat{\beta}_1^*$ and $\hat{\beta}_0 = \hat{\beta}_0^*/(1 - \hat{\rho})$ as the estimates in (6.177);
- (v) if the residuals from fitting the transformed model show no autocorrelation then stop. Otherwise repeat the procedure starting with the estimated transformed data;
- (vi) iterate to convergence.

It is often recommended that only the first step be used [27].

Example 6.13 To demonstrate the use of the Durbin-Watson statistic, we refer to the Longley data fit with the two best predictors x_2 and x_3 . The R^2 for this model is 0.981, and all three regression coefficients are highly significant. On the surface it would seem that the analysis looks complete and that a good-fitting model has been found. However, clearly as we see in Figure 6.23, a plot of the residuals against the fitted values strongly suggests the presence of autocorrelation in the residuals. In addition to graphical displays, in order to use the Durbin-Watson procedure to detect the autocorrelated errors, the computations for d and $\hat{\rho}$ are displayed in Table 6.20. Recall that we assume the residuals follow a first-order autoregressive model.

From Table 6.20 and using (6.173) we obtain the value of the Durbin-Watson statistic

$$d = \frac{3493839}{3579065} = 0.9762,$$

and using (6.174) the estimate of the autocorrelation parameter $\hat{\rho}$ is

$$\hat{\rho} = \frac{1557272}{3155999} = 0.4934.$$

From Table A.4 in Appendix, for $\alpha = 5\%$, $n = 16$, and $m = 2$, we observe $d_L = 0.98$. Since $d < d_L$, we reject the hypothesis $H_0 : \rho = 0$ and conclude that $\rho > 0$. If d is

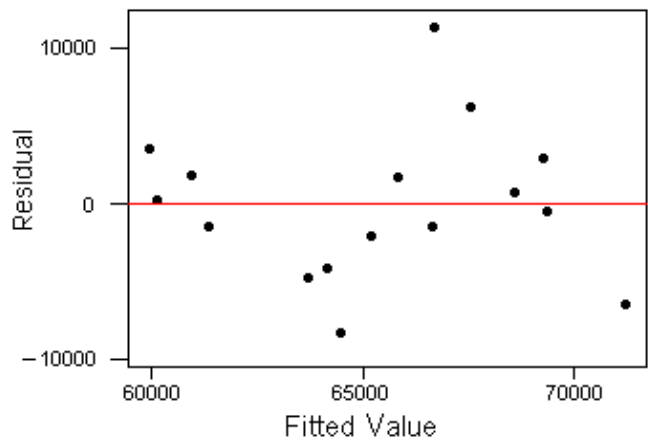


Figure 6.23: Plot of residuals versus fitted values

significant, we go through the Cochrane-Orcutt procedure using the transformed variables to remove the autocorrelation. We leave this to the reader for an exercise.

Table 6.20 Computing the Durbin-Watson Statistic

No.	$\hat{\varepsilon}_i$	$\hat{\varepsilon}_{i-1}$	$\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}$	$(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2$	$\hat{\varepsilon}_i^2$	$\hat{\varepsilon}_{i-1}^2$	$\hat{\varepsilon}_i \hat{\varepsilon}_{i-1}$
1	355.92	-	-	-	126680	-	-
2	186.88	355.92	-169.043	28576	34924	126680	66514
3	25.43	186.88	-161.453	26067	646	34924	4752
4	-142.97	25.43	-168.395	28357	20440	646	-3635
5	-468.73	-142.97	-325.757	106118	219704	20440	67013
6	-823.54	-468.73	-354.811	125891	678213	219704	386013
7	-202.97	-823.54	620.566	385102	41197	678213	167154
8	-416.53	-202.97	-213.564	45610	173501	41197	84544
9	175.02	-416.53	591.551	349932	30631	173501	-72900
10	1146.89	175.02	971.876	944542	1315360	30631	200724
11	628.24	1146.89	-518.648	268996	394690	1315360	720527
12	-146.46	628.24	-774.705	600168	21451	394690	-92014
13	79.80	-146.46	226.266	51196	6369	21451	-11688
14	300.04	79.80	220.233	48502	90022	6369	23944
15	-46.58	300.04	-346.622	120147	2170	90022	-13977
16	-650.44	-46.58	-603.851	364635	423066	2170	30300
Sum				3493839	3579065	3155999	1557272

The Hildreth-Lu procedure is a modification of the Cochrane-Orcutt procedure in that it seeks to minimize the SSE of the transformed model (6.121) simultaneously with ρ . That is, we find $(\beta_0^*, \beta_1^*, \hat{\rho})$ to minimize

$$\sum_{t=2}^n (y_t^* - \beta_0^* - \beta_1^* x_t^*)^2.$$

(6.181)

Generally, this requires more work than the Cochrane-Orcutt method since minimizing (6.181) is a nonlinear optimization problem. On the other hand, the Cochrane-Orcutt procedure can be carried out using a straightforward modification of a standard OLS program.

Last, we note that autocorrelation may appear for reasons other than some inherent property of the data.

6.7 Generalized Least Squares

As a last topic in this chapter we consider the problem of least squares estimation when the errors ϵ in (5.1) have an arbitrary variance-covariance matrix \mathbf{W} . If \mathbf{W} is known, then least squares estimation can be done, as for weighted regression, by converting the model to an equivalent one with uncorrelated errors with constant variance.

Assume now that

$$Y_i = \beta_0 + \sum_{j=1}^n x_{ij}\beta_j + \epsilon_i, \quad 1 \leq i \leq n \quad (6.182)$$

when $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ has a joint multivariate normal distribution with $E(\epsilon) = 0$ and $\Sigma(\epsilon) = \mathbf{W}$. Then, the joint likelihood density function is given by

$$f_y(y) = \left(\sqrt{2\pi} |\mathbf{W}|\right)^{-n/2} \exp \left[-\langle y - \mu, \mathbf{W}^{-1}(y - \mu) \rangle / 2 \right] \quad (6.183)$$

where $\mu = \mathbf{X}\beta$.

Thus the MLE of β is found by minimizing the positive definite quadratic form

$$\mathbf{Q} = \langle y - \mu, \mathbf{W}^{-1}(y - \mu) \rangle. \quad (6.184)$$

Since \mathbf{W} is symmetric it can be factored as

$$\mathbf{W} = \mathbf{R}\mathbf{R}^T \quad (6.185)$$

and

$$\mathbf{W}^{-1} = \left(\mathbf{R}\mathbf{R}^T\right)^{-1} = \left(\mathbf{R}^T\right)^{-1} \mathbf{R}^{-1} = \left(\mathbf{R}^{-1}\right)^T \mathbf{R}^{-1} \quad (6.186)$$

where \mathbf{R} is nonsingular (this can be done, for example as in Theorem 4.12 or using Cholesky factorization). Then,

$$\begin{aligned} \mathbf{Q} &= \langle y - \mu, (\mathbf{R}^{-1})^T \mathbf{R}^{-1}(y - \mu) \rangle \\ &= \langle \mathbf{R}^{-1}(y - \mu), \mathbf{R}^{-1}(y - \mu) \rangle \\ &= \langle \mathbf{R}^{-1}y - \mathbf{R}^{-1}\mathbf{X}\beta, \mathbf{R}^{-1}y - \mathbf{R}^{-1}\mathbf{X}\beta \rangle \\ &= \langle z - \mathbf{X}_R\beta, z - \mathbf{X}_R\beta \rangle \end{aligned} \quad (6.187)$$

where $z = \mathbf{R}^{-1}y$ and $\mathbf{X}_R = \mathbf{R}^{-1}\mathbf{X}$.

Now \mathbf{Q} is in the form occurring in (5.83) with \mathbf{z} replacing \mathbf{y} and \mathbf{X}_R replacing \mathbf{X} . Hence it follows from Theorem 5.1 that \mathbf{Q} is minimized by choosing

$$\begin{aligned}\hat{\beta}_{\mathbf{W}} &= (\mathbf{X}_R^T \mathbf{X}_R)^{-1} \mathbf{X}_R^T \mathbf{z} = \left[(\mathbf{R}^{-1} \mathbf{X})^T \mathbf{R}^{-1} \mathbf{X} \right]^{-1} (\mathbf{R}^{-1} \mathbf{X})^T \mathbf{R}^{-1} \mathbf{y} \\ &= \left[\mathbf{X}^T (\mathbf{R}^{-1})^T \mathbf{R}^{-1} \mathbf{X} \right]^{-1} \mathbf{X}^T (\mathbf{R}^{-1})^T \mathbf{R}^{-1} \mathbf{y} \\ &= \left[\mathbf{X}^T (\mathbf{R} \mathbf{R}^T)^{-1} \mathbf{X} \right]^{-1} \mathbf{X}^T (\mathbf{R} \mathbf{R}^T)^{-1} \mathbf{y} \\ &= (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{y}.\end{aligned}\tag{6.188}$$

$\hat{\beta}_{\mathbf{W}}$ is called the *generalized least squares estimator* of β .

When the errors are not normal, then this argument suggests that we estimate β by $\hat{\beta}_{\mathbf{W}}$ in this case as well. In fact, it can be shown, generalizing the Gauss-Markov theorem, that $\hat{\beta}_{\mathbf{W}}$ is the BLUE estimator of β in (6.182). When $\mathbf{W} = \sigma^2 \mathbf{I}_n$ $\hat{\beta}_{\mathbf{W}}$ reduces to $\hat{\beta}_{OLS}$ and when $\mathbf{W} = \sigma^2 \text{diag}(1/w_1, 1/w_2, \dots, 1/w_n)$ it is the weighted least squares estimator $\hat{\beta}_{\mathbf{W}}$.

For further discussion it is useful to consider to $\hat{\beta}_{\mathbf{W}}$ as the OLS estimator for the transformed model

$$\mathbf{Z} = \mathbf{X}_R \beta + \delta \tag{6.189}$$

where $\mathbf{Z} = \mathbf{R}^{-1} \mathbf{Y}$ and $\delta = \mathbf{R}^{-1} \epsilon$. Using (6.189)

$$\begin{aligned}\Sigma(\delta) &= \mathbf{R}^{-1} \Sigma(\epsilon) (\mathbf{R}^{-1})^T = \mathbf{R}^{-1} \mathbf{W} (\mathbf{R}^{-1})^T \\ &= \mathbf{R}^{-1} \mathbf{R} \mathbf{R}^T (\mathbf{R}^{-1})^T = \mathbf{I}_n\end{aligned}\tag{6.190}$$

so that the errors in (6.182) are uncorrelated with $\text{Var}(\delta_i) = 1, 1 \leq i \leq n$. Using this we can easily establish the BLUE property of $\hat{\beta}_{\mathbf{W}}$ and derive a number of other useful properties of $\hat{\beta}_{\mathbf{W}}$. We summarize these in Theorem 6.5.

Theorem 6.5 *Consider the GLM (6.182) where $\Sigma(\epsilon) = \mathbf{W}$ is nonsingular. Then,*

- (i) *If the errors are $\mathbf{N}(\mathbf{0}, \mathbf{W})$ then the generalized least squares estimator $\hat{\beta}_{\mathbf{W}}$ is the best estimator of β .*
- (ii) *If the errors have an arbitrary joint distribution with $E(\epsilon) = 0$, then $\hat{\beta}_{\mathbf{W}}$ is the BLUE estimator of β .*
- (iii) *If $\Sigma(\epsilon) = \sigma^2 \mathbf{I}_n$, then $\hat{\beta}_{\mathbf{W}} = \hat{\beta}_{OLS}$.*
- (iv) *The variance-covariance matrix of $\hat{\beta}_{\mathbf{W}}$, $\Sigma(\hat{\beta}_{\mathbf{W}})$ is given by*

$$\Sigma(\hat{\beta}_{\mathbf{W}}) = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1}.\tag{6.191}$$

Proof. (i) This can be established using the fact that $\hat{\beta}_{OLS}$ is the standard GLM estimator of β and the relation of (6.182) to the transformed model. For details see [27, 85].

(ii) Again this can be proven using the transformed model. Details are left to the reader.

(iii) This has been established above.

For (iv) we have

$$\begin{aligned}
 \Sigma(\hat{\beta}_{\mathbf{W}}) &= (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \Sigma(\mathbf{Y}) \mathbf{W}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \\
 &= (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{W} \mathbf{W}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \\
 &= (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}) (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \\
 &= (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1}.
 \end{aligned} \tag{6.192}$$

Since $\hat{\beta}_{\mathbf{W}}$ is unbiased, $\hat{\beta}_{\mathbf{W}}$ is $\mathbf{N}(\beta, (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1})$ if ε is $\mathbf{N}(\mathbf{0}, \mathbf{W})$. ■

For purposes of calculation and testing it is usually more convenient to use the transformation $\mathbf{Z} = \mathbf{R}^{-1} \mathbf{Y}$, $\mathbf{X}_R = \mathbf{R}^{-1} \mathbf{X}$ in going from (6.189) to (6.190) rather than computing $\hat{\beta}_{\mathbf{W}}$ directly. Doing this we arrive at the following algorithm for carrying out the least squares analysis of the model in (6.182).

(i) Factor \mathbf{W} as $\mathbf{R}\mathbf{R}^T$. (Say using Cholesky factorization.)

(ii) Transform the data (\mathbf{y}, \mathbf{X}) to $(\mathbf{R}^{-1} \mathbf{y}, \mathbf{R}^{-1} \mathbf{X}) = (\mathbf{z}, \mathbf{X}_R)$.

(iii) Regress \mathbf{z} on the columns of \mathbf{X}_R using OLS to obtain $\hat{\beta}_{\mathbf{W}}$.

(iv) Since $\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} = \mathbf{X}^T (\mathbf{R}\mathbf{R}^T)^{-1} \mathbf{X} = \mathbf{X}^T (\mathbf{R}^{-1})^T \mathbf{R}^{-1} \mathbf{X} = \mathbf{X}_R^T \mathbf{X}_R$, $\Sigma(\hat{\beta}_{\mathbf{W}}) = (\mathbf{X}_R^T \mathbf{X}_R)^{-1}$ so that $\text{Var}(\hat{\beta}_{\mathbf{W},i})$ is the i -th diagonal element δ_i of $(\mathbf{X}_R^T \mathbf{X}_R)^{-1}$.

(v) When the errors are normal $(\hat{\beta}_{\mathbf{W},i} - \beta_i) / \sqrt{\delta_i}$ is $N(0, 1)$ so that a $(1 - \alpha) \times 100\%$ confidence intervals for β_i are obtained from

$$\hat{\beta}_{\mathbf{W},i} \pm z_{\alpha/2} \sqrt{\delta_i}, \quad 0 \leq i \leq m \tag{6.193}$$

and these may be used in the usual way to conduct hypothesis tests concerning the coefficients β_i .

(vi) Tests of the general linear hypothesis $\mathbf{C}\beta = \mathbf{b}$ when $\varepsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{W})$ can be carried out using the “extra sum of squares principle” on the transformed model (6.121). In this case, since $\sigma^2 = 1$ is known, ΔSSE will have a $\chi^2(r)$ distribution when H_0 is true, where $r = \text{rank}(\mathbf{C})$. Thus $H_0 : \mathbf{C}\beta = \mathbf{b}$ is rejected at level α if $\Delta SSE > \chi^2_{r,\alpha}$.

In particular, one can test for the overall significance of the regression by fitting the reduced model

$$z_i = x_{R_{i0}} \beta_0 + \delta_i, \quad 1 \leq i \leq n \tag{6.194}$$

where $x_{R_{i0}}$ is the i o-th element of \mathbf{X}_R , forming the residual sum of squares and subtracting the SSE from fitting the full model (6.182). The hypothesis $H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0$ is rejected at level α if $\Delta SSE > \chi^2_{m,\alpha}$.

Since the transformed model (6.121) does not have an intercept, the same problem arises as to an appropriate goodness of fit measure as for WLS. A choice consistent with those given previously is to use the squared sample correlation coefficient of \mathbf{z} and $\hat{\mathbf{z}} = \mathbf{X}_R \hat{\boldsymbol{\beta}}_{\mathbf{W}}$ in (6.189). Another choice, given by Buse can be defined as follows.

Let

$$\bar{y}_{\mathbf{W}} = \langle \mathbf{1}, \mathbf{W}^{-1} \mathbf{y} \rangle / \langle \mathbf{n}, \mathbf{W}^{-1} \mathbf{n} \rangle \quad (6.195)$$

where $\mathbf{1} = (1, 1, \dots, 1)^T$, $\mathbf{n} = n\mathbf{1}$ and $\bar{y}_{\mathbf{W}}$ denote the *weighted mean* of the observations y_i , $1 \leq i \leq n$. (It is easily shown that $\bar{y}_{\mathbf{W}}$ is the generalized least squares estimator in the reduced model $Y_i = \beta_0 + \varepsilon_i$, $1 \leq i \leq n$, $\boldsymbol{\Sigma}(\boldsymbol{\varepsilon}) = \mathbf{W}$.) Then it can be shown that the following decomposition holds: letting $\hat{\mathbf{Y}} = \mathbf{X} \hat{\boldsymbol{\beta}}_{\mathbf{W}}$

$$\underbrace{\langle \mathbf{y} - \bar{y}_{\mathbf{W}} \mathbf{1}, \mathbf{W}^{-1} (\mathbf{y} - \bar{y}_{\mathbf{W}} \mathbf{1}) \rangle}_{SST} = \underbrace{\langle \hat{\mathbf{y}} - \bar{y}_{\mathbf{W}} \mathbf{1}, \mathbf{W}^{-1} (\hat{\mathbf{y}} - \bar{y}_{\mathbf{W}} \mathbf{1}) \rangle}_{SSR} + \underbrace{\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{W}^{-1} (\mathbf{y} - \hat{\mathbf{y}}) \rangle}_{SSE}. \quad (6.196)$$

Since each of the quadratic forms in (6.196) is positive definite, we can define

$$R_{\mathbf{W}}^2 = SSR/SSR \quad (6.197)$$

which has properties analogous to the usual R^2 and reduces to it when $\mathbf{W} = \sigma^2 \mathbf{I}_n$.

When \mathbf{W} is unknown, as we have already seen for WLS, the problem of estimating $\boldsymbol{\beta}$ is generally intractable unless further assumptions are made concerning the error structure. If \mathbf{W} can be parameterized by a small number of parameters, then maximum likelihood estimation may be possible as well as generalizations of iteratively re-weighted least squares. These topics are dealt with at length in the literature on time series analysis [11] and kriging [23] and will not be further dealt with in this text.

6.8 Exercises

- 6.1** [90] An experiment was conducted to gain some preliminary insight into the effect of three quantitative factors on the capability of a particular coal-cleaving operation. A polymer was used to clean the coal and the amount, y was measured (mg/l). The factors that influenced the suspended solids are

- x_1 : percentage in solids in the input solution
- x_2 : pH of the tank that holds the solution
- x_3 : flow rate of the cleaving polymer, ml/minute

Assume that all three factors were controlled in the experimental process. The data are given in Table 6.21.

For the multiple regression model containing all three regressors, calculate the following influence diagnostics.

- (a) Residuals $\hat{\varepsilon}_i$ and \hat{t}_i , and make plots against fitted values.
- (b) Leverages - HAT diagonals h_{ii} .
- (c) Cook's D_i .
- (d) DFFITS $_i$, and the cut-off point.
- (e) DFBETAS $_{j,i}$ ($j = 1, 2, 3$), and the cut-off point.

Table 6.21 Coal-cleaving Data

No.	x_1	x_2	x_3	y
1	1.5	6.0	1315	243
2	1.5	6.0	1315	261
3	1.5	9.0	1890	244
4	1.5	9.0	1890	285
5	2.0	7.5	1575	202
6	2.0	7.5	1575	180
7	2.0	7.5	1575	183
8	2.0	7.5	1575	207
9	2.5	9.0	1315	216
10	2.5	9.0	1315	160
11	2.5	6.0	1890	104
12	2.5	6.0	1890	110

6.2 Show that $\hat{\beta}_{\mathbf{W}}$, the generalized least squares estimator given by (6.188), is unbiased for $\beta_{\mathbf{W}}$.

6.3 Show that the vector $\hat{\beta}_{\mathbf{W}}$, which minimizes

$$(\mathbf{y} - \mathbf{X}\beta_{\mathbf{W}})^T \mathbf{W}^{-1} (\mathbf{y} - \mathbf{X}\beta_{\mathbf{W}})$$

is given by the generalized least squares estimator given in (6.188).

6.4 An experiment was conducted to study the growth characteristics of corn roots in the presence of a particular herbicide that was applied to a certain type of soil. Data were collected and percent of control, meaning the percent of the growth observed without the herbicide, was used as the response [90].

Obs. No.	Concentration of Herbicide (x)	Percentage of Control (y)
1	0.5	95.8467
2	1.0	91.6561
3	2.0	81.5142
4	8.0	75.7477
5	32.0	68.7061
6	128.0	35.9895

(a) Plot the data and give a comment.

(b) Using the transformation $\log(y_i)$, fit the data to the model

$$\log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

(c) Find s^2 for the model in (b).

(d) Compute the original residuals, the PRESS residuals, and the sum of the absolute PRESS residuals.

6.5 For the regression models shown below, determine whether it is a linear model, an intrinsically linear model, or a intrinsically nonlinear model. If the model is intrinsically linear, suggest how it can be linearized by a suitable transformation.

(a) $y = \beta_1 e^{\beta_2 x} \varepsilon$

(b) $y = \beta_1 \exp(\beta_2 + \beta_3 x) + \varepsilon$

(c) $y = \beta_1 + \beta_2 \exp(\beta_3 x) + \varepsilon$

(d) $y = \theta_1 + (\theta_2/\theta_1)x + \varepsilon$

(e) $y = \theta_1 + \theta_2 x_1 + \theta_2 \left(x_2^{\theta_3} \right) + \varepsilon$

6.6 Using two models in (6.108) and (6.109) for tree data in Table 6.7, make plots of the studentized residuals \hat{r}_i against the fitted value D and H respectively. Do you observe any outlier(s)?

6.7 What happens to the weighted average

$$\bar{x}_w = \frac{w_1^2 x_1 + w_2^2 x_2}{w_1^2 + w_2^2} \quad (6.198)$$

if the first weight w_1 approaches zero? The measurement x_1 is totally unreliable.

6.8 Suppose that you have n independent measurements x_1, x_2, \dots, x_n from your pulse rate, weighted by w_1, w_2, \dots, w_n , what is the weighted average that replaces (6.198)? It is the best estimate when the statistical variances are $\sigma_i^2 = 1/w_i^2$.

6.9 (Cauchy-Schwarz inequality) If X and Y have means μ_X, μ_Y and variances σ_X^2, σ_Y^2 , respectively, prove

$$[E(XY)]^2 \leq E(X^2) E(Y^2).$$

6.10 Consider the first-order autoregressive model for a sample of size $n = 32$:

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

where $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ with $|\rho| < 1$ and u_t are independent $N(0, \sigma^2)$.

(a) Explain the procedure for testing $H_0 : \rho = 0$ versus $H_1 : \rho > 0$ at $\alpha = 0.05$.

(b) Explain the procedure for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ at $\alpha = 0.01$.

6.11 A study was conducted by McNamara and Browne [80] to examine the relationship between the prices of gold and “black gold” over a period of time. Table 6.22 presents quarterly data from March 1976 to March 1980 on the price of gold (\$/ounce) and the price of petroleum (\$/barrel). Take the price of gold as a response variable Y and the price of gold an explanatory variable x .

(a) Fit a simple linear regression model to these data.

(b) Determine whether or not the assumption of independence of residuals in the OLS model appears to have been violated.

(1) by plotting the residuals against time.

(2) by calculating Durbin-Watson test.

(c) If necessary, correct the autocorrelation using the Cochrane-Orcutt method.

Table 6.22 Petroleum and Gold Prices in United States

Year	Month	Petroleum	Gold	Year	Month	Petroleum	Gold
1976	Mar.	7.79	133.1	1978	Mar.	8.80	184.1
1976	Jun.	7.99	126.2	1978	Jun.	9.05	184.1
1976	Sep.	8.39	114.7	1978	Sep.	9.12	212.4
1976	Dec.	8.55	134.4	1978	Dec.	9.27	208.1
1977	Mar.	8.45	148.6	1979	Mar.	9.83	242.4
1977	Jun.	8.44	140.8	1979	Jun.	11.70	279.4
1977	Sep.	8.63	150.1	1979	Sep.	14.57	357.2
1977	Dec.	8.75	161.1	1979	Dec.	17.03	459.0
				1980	Mar.	19.35	553.6

6.12 Find the weighted least squares solution $\bar{\mathbf{x}}_{\mathbf{W}}$ to $\mathbf{Ax} = \mathbf{b}$:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Check that the projection $\mathbf{A}\bar{\mathbf{x}}_{\mathbf{W}}$ is still perpendicular (in the \mathbf{W} -inner product!) to the error $\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}_{\mathbf{W}}$.

6.13 Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the ε_i 's are independent with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = x_i^2 \sigma^2$, $i = 1, 2, \dots, n$.

(a) Show that weighted least squares estimation (WLS) is equivalent to ordinary least squares estimation (OLS) for the model [104]

$$\left(\frac{Y_i}{x_i} \right) = \beta_1 + \beta_0 \left(\frac{1}{x_i} \right) + \delta_i.$$

(b) Under the normality assumption of errors δ_i in the model in (a), find the MLEs of β_1 , β_0 .

6.14 Use the Cochrane-Orcutt procedure to correct for the autocorrelation in the Longley data in Example 6.3.

Chapter 7

Further Applications of Regression Techniques

7.1 Introduction

In this chapter we will expand on a number of regression modelling techniques that were discussed briefly in Chapters 5 and 6. These will include a further discussion of polynomial and piecewise polynomial models in one and several variables, the further use of dummy (indicator) variables to deal with qualitative factors in modelling and last a further discussion of binary response models and logistic regression. These techniques allow one to model a wide variety of phenomena in science and technology.

7.2 Polynomial Models in One Variable

As we indicated in Chapter 5, by choosing $x_j = x^j, 0 \leq j \leq m$, in the GLM the model (5.1) becomes

$$Y = \beta_0 + \sum_{j=1}^m \beta_j x^j + \varepsilon \quad (7.1)$$

which we refer to as a *polynomial model of degree m* . Here $E(Y)$ depends *nonlinearly* on x but is a *linear model* since it depends *linearly* on the parameters $\beta_j, 0 \leq j \leq m$.

Generally a polynomial model is suggested if a scatter plot of (x, y) shows substantial curvature or a residual plot from a linear fit in x shows curvature as well. When the behavior in x appears monotone, then variable transformations as discussed in Chapters 3 and 6 are often appropriate. However, if the curve is not monotone such as that shown in Figure 7.1, then a polynomial model may be more appropriate. For modelling over large ranges of x , *piecewise polynomial (spline) models* have many advantages. These will be discussed in the following section.

Although fitting a polynomial appears to be a straightforward application of the GLM, there are a number of pitfalls that one should be aware of in using a polynomial model. First the normal equations arising from (7.1) can be highly ill-conditioned, even for low order polynomials, particularly if the observations are taken over a small range of x . This can lead to substantial round-off errors which one should be aware of when

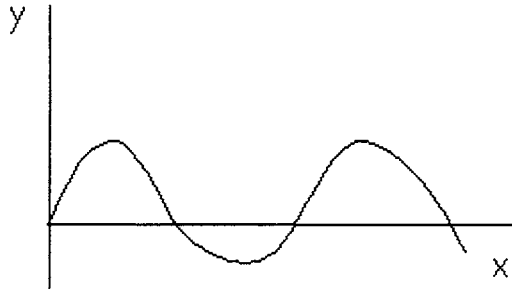


Figure 7.1: A non-monotonic curve

using “off-the-shelf” software. Fortunately, there are a number of standard certified data sets which are available (they can be found on the Internet) to test the accuracy of one’s software.

Second, there is the issue of choosing the degree m of the polynomial in (7.1). Since a polynomial of degree m can be fitted exactly to $m + 1$ data points, one needs to be cautious of *overfitting* the data. Generally, one should use as low order polynomial as possible to obtain a satisfactory fit. Models which are overfit will not be useful for prediction.

To deal with the ill-conditioning, a widely recommended remedy is to use centered variables

$$z^j = (x - \bar{x})^j \quad (7.2)$$

where \bar{x} is the mean of the independent variable. One then fits the model

$$Y = \beta_0 + \sum_{j=1}^m \gamma_j z^j + \varepsilon \quad (7.3)$$

rather than (7.1). By expanding z^j using the binomial theorem one can then relate γ_j to β_j , $0 \leq j \leq m$, in (7.1). Another, more sophisticated approach is to use orthogonal polynomials which will be discussed in this Section.

Choosing the degree of the polynomial is a more difficult problem, characteristic of all model building. Too low a degree may give a poor fit, too high a degree may give a good fit, but be unreliable for prediction. Anticipating our discussion of subset selection in Chapter 8, we examine two approaches - forward selection and backward elimination.

In forward selection, one can start with a linear (in x) fit and examine the summary statistics and residual plots. If γ_j is significant, R^2 is large and residual plots show no evidence of curvature, we might stop. If either R^2 is small and residual plots of $\hat{\varepsilon}$ against x or against \hat{y} show curvature, then one might add a quadratic term. This process can be repeated, until an adequate fit is obtained with significant coefficients, large R^2 and appropriate residuals.

This approach has its pitfalls, because one might stop too soon. For example, if the true model is

$$Y = \beta_0 + \beta_1 x + \beta_3 x^3 + \varepsilon \quad (7.4)$$

then, the procedure outlined above might suggest that a quadratic term is not present, since a t test might indicate that $\beta_2 = 0$ and one might stop adding variables, even though the model contains a cubic term. Of course, if R^2 is still small and residual plots show distinct curvature one might consider higher order terms.

In backward elimination one starts with a polynomial of sufficiently high degree less than the number of data points and deletes terms whose coefficients have small t values. Again, because of multicollinearity one must be careful in just eliminating variables "en-masse". Elimination of one variable at a time is generally a preferred approach. If one uses orthogonal polynomials, then these can be entered in any order and provide a less ambiguous choice of model.

7.2.1 Orthogonal Polynomials

As we have noted, even though the ill-conditioning can be alleviated by centering, there may still exist a high level of multicollinearity which usually causes computational difficulties. These difficulties can be avoided using orthogonal polynomials. Orthogonal polynomials are uncorrelated.

Suppose that the model is given in (7.1). Since the columns of the design matrix \mathbf{X} will not be orthogonal, we now consider the model

$$Y_i = \gamma_0 \psi_0(x_i) + \gamma_1 \psi_1(x_i) + \cdots + \gamma_m \psi_m(x_i) + \varepsilon_i \quad (7.5)$$

where $\psi_r(x_i)$ is an r -th degree polynomial ($r = 1, 2, \dots, m$) in the x_i 's ($i = 1, 2, \dots, n$) such that

$$\sum_{i=1}^n \psi_r(x_i) \psi_s(x_i) = 0, \text{ for all } r, s, r \neq s \quad (7.6)$$

and

$$\psi_0(x_i) = 1, i = 1, 2, \dots, n. \quad (7.7)$$

Then the model in (7.5) becomes $\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where

$$\mathbf{X} = \begin{bmatrix} \psi_0(x_1) & \psi_1(x_1) & \cdots & \psi_m(x_1) \\ \psi_0(x_2) & \psi_1(x_2) & \cdots & \psi_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_0(x_n) & \psi_1(x_n) & \cdots & \psi_m(x_n) \end{bmatrix}.$$

Since the columns of the \mathbf{X} matrix are orthogonal,

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} \sum_{i=1}^n \psi_0^2(x_i) & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n \psi_1^2(x_i) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^n \psi_m^2(x_i) \end{bmatrix} \quad (7.8)$$

and the least square estimator is given by $\hat{\boldsymbol{\gamma}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, we have

$$\hat{\gamma}_j = \frac{\sum_{i=1}^n \psi_j(x_i) Y_i}{\sum_{i=1}^n \psi_j^2(x_i)}, j = 0, 1, 2, \dots, m. \quad (7.9)$$

Since $\psi_0(x_i) = 1$, it follows from (7.9) that

$$\hat{\gamma}_0 = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}. \quad (7.10)$$

The residual sum of squares is then

$$\begin{aligned} SSE &= (\mathbf{Y} - \mathbf{X}\hat{\gamma})^T (\mathbf{Y} - \mathbf{X}\hat{\gamma}) = \mathbf{Y}^T \mathbf{Y} - \hat{\gamma}^T \mathbf{X}^T \mathbf{X} \hat{\gamma} \\ &= \sum_{i=1}^n Y_i^2 - \sum_{j=0}^m \left[\sum_{i=1}^n \psi_j^2(x_i) \right] \hat{\gamma}_j^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{j=1}^m \left[\sum_{i=1}^n \psi_j^2(x_i) \right] \hat{\gamma}_j^2. \end{aligned} \quad (7.11)$$

If one wishes to test $H_0 : \gamma_m = 0$, which is in fact equivalent to testing $H_0 : \beta_m = 0$ in (7.1), the residual sum of squares under the null hypothesis is

$$\begin{aligned} SSE_{H_0} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{j=1}^{m-1} \left[\sum_{i=1}^n \psi_j^2(x_i) \right] \hat{\gamma}_j^2 \\ &= SSE + \hat{\gamma}_m^2 \sum_{i=1}^n \psi_m^2(x_i). \end{aligned} \quad (7.12)$$

Then, the test statistic would be

$$F = \frac{SSE_{H_0} - SSE}{SSE/(n - m - 1)} = \frac{\hat{\gamma}_m^2 \sum_{i=1}^n \psi_m^2(x_i)}{SSE/(n - m - 1)}. \quad (7.13)$$

The orthogonal polynomials $\psi_j(x_i)$ can be obtained in many different ways. In particular, if the levels of x are equally spaced, they can be easily constructed. A survey of methods for generating orthogonal polynomials can be found in Seber [104]. We note that the method of generating the ψ_r is similar to Gram-Schmit orthogonalization, with the difference that only the preceding two polynomials are involved at each stage.

Some of these orthogonal polynomials are given in Table 7.1.

Table 7.1 Coefficients of Orthogonal Polynomials

x_j	$n = 3$		$n = 4$			$n = 5$				$n = 6$				
	ψ_1	ψ_2	ψ_1	ψ_2	ψ_3	ψ_1	ψ_2	ψ_3	ψ_4	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
1	-1	1	-3	1	-1	-2	2	-1	1	-5	5	-5	1	-1
2	0	-2	-1	-1	3	-1	-1	2	-4	-3	-1	7	-3	5
3	1	1	1	-1	-3	0	-2	0	6	-1	-4	4	2	-10
4			3	1	1	1	-1	-2	-4	1	-4	-4	2	10
5						2	2	1	1	3	-1	-7	-3	-5
6										5	5	5	1	1
$\sum_{j=1}^n \psi_j^2$	2	6	20	4	20	10	14	10	70	70	84	180	28	252

For the case $n \geq 7$ readers should refer to [24, 94, 27].

7.2.2 Piecewise Polynomial Models

Although polynomial models are conceptually quite simple as a way of dealing with non-linear trends in a predictor variable x , there are some difficulties, particularly if the range of the data is large. In this case a piecewise polynomial model can be computationally and conceptually easier to use. Such problems often occur in economic data where the independent variable represents time and various time trends may be present.

Example 7.1 Consider the data in Table 7.2 below.

Table 7.2 A Data									
x	1	2	3	4	5	6	7	8	9
y	2.3	2.8	6.5	7.4	10.2	10.5	12.1	13.2	13.6

Graphing this data in Figure 7.2 indicates that the first five data points lie on one line while the final four appear to lie on another.

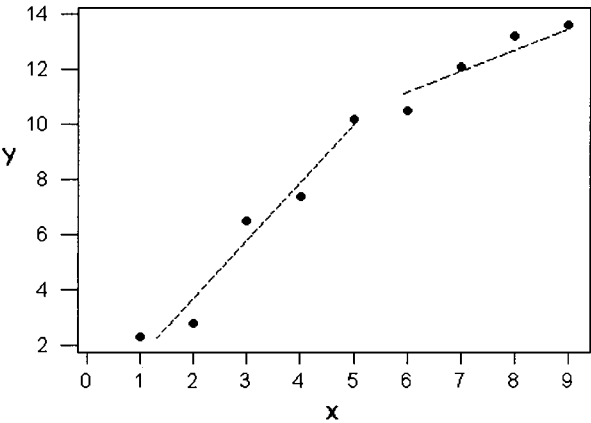


Figure 7.2: Scatter plot of data in Example 7.1

If this is the case, then we can perform two regressions - one for the first five points and another for the last four and be done. However, one of the things we might like to know is whether our visual impression is correct. For example, do the slopes of the two lines really differ? Another reasonable question to ask is whether $x = 5$ is the abscissa of the point of intersection of the two lines. That is, is the true model of the form in Figure 7.3 or Figure 7.4. As we shall see, an appropriate choice of multiple regression models allows one to make such inferences in a straightforward way.

Historically, such problems appear to have been treated using dummy variables, such as the approach given in Draper and Smith [27]. More recently the use of *spline functions* has been advocated [122, 106, 117] and this is the approach we follow.

Consider, for some number x_0 , the two functions

$$(x - x_0)_+^0 = \begin{cases} 0, & x \leq x_0, \\ 1, & x > x_0, \end{cases} \tag{7.14}$$

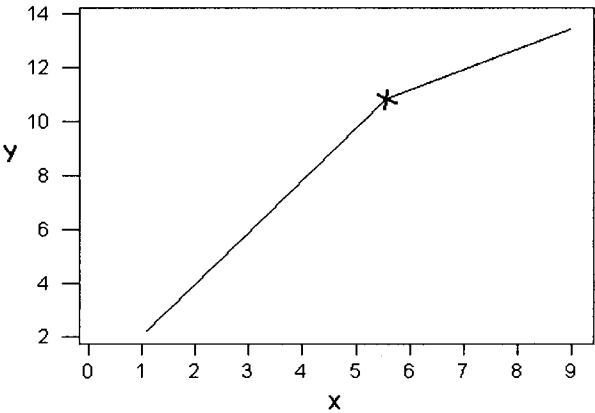


Figure 7.3: The true model A of the data

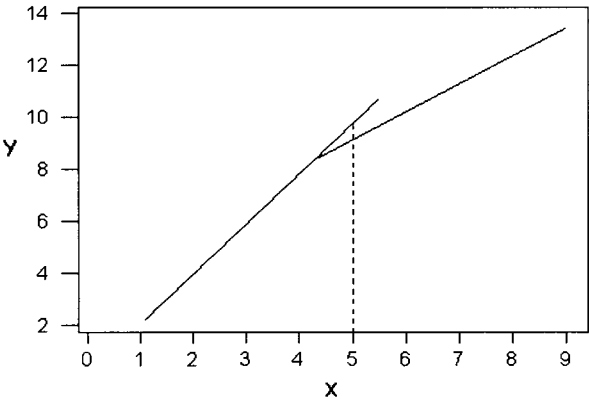


Figure 7.4: The true model B of the data

and

$$(x - x_0)_+ = \begin{cases} 0, & x \leq x_0, \\ x - x_0, & x > x_0. \end{cases} \quad (7.15)$$

To model two straight lines simultaneously, we consider the linear model

$$E(Y) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)_+^0 + \beta_3(x - x_0)_+. \quad (7.16)$$

To see that (7.16) actually represents two lines, we consider $x \leq x_0$ and $x > x_0$ separately. Now if $x \leq x_0$, then,

$$E(Y) = \beta_0 + \beta_1(x - x_0) = \beta_0 - \beta_1 x_0 + \beta_1 x \quad (7.17)$$

and this is a straight line with slope β_1 and intercept $\beta_0 - \beta_1 x_0$. If $x > x_0$, then,

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1(x - x_0) + \beta_2 + \beta_3(x - x_0) \\ &= \beta_0 - \beta_1 x_0 - \beta_3 x_0 + \beta_2 + (\beta_1 + \beta_3)x \end{aligned} \quad (7.18)$$

and this is a line with slope $\beta_1 + \beta_3$ and intercept $\beta_0 - \beta_1 x_0 - \beta_3 x_0 + \beta_2$. We also note that at $x = x_0$ the value of $E(Y)$ on the first line is β_0 while the value of $E(Y)$ on the second line is $\beta_0 + \beta_2$. Thus, we can test whether the slopes are equal by testing whether $\beta_3 = 0$, while the assumption that x_0 is the abscissa of the point of intersection may be checked by testing whether $\beta_2 = 0$. (We note that β_2 is the difference in the vertical heights of the lines at $x = x_0$.) To illustrate some of these ideas numerically we consider the data given in Table 7.3.

In this case the proposed model is

$$Y = \beta_0 + \beta_1(x - 5) + \beta_2(x - 5)_+^0 + \beta_3(x - 5)_+ + \varepsilon. \quad (7.19)$$

Using (7.14) and (7.15) we arrive at the following table of data to perform the regression.

Table 7.3 Some data

y	x_0	x_1	x_2	x_3
2.3	1	-4	0	0
3.8	1	-3	0	0
6.5	1	-2	0	0
7.4	1	-1	0	0
10.2	1	0	1	0
10.5	1	1	1	1
12.1	1	2	1	2
13.2	1	3	1	3
13.6	1	4	1	4

where $x_1 = (x - 5)$, $x_2 = (x - 5)_+^0$ and $x_3 = (x - 5)_+$. Fitting this data by least squares we obtain

$$\hat{y} = 9.2 + 1.94(x - 5) - 0.17(x - 5)_+^0 - 0.9(x - 5)_+ \quad (7.20)$$

as the estimated model.

The t values are: $t_0 = 27.33$; $t_1 = 13.09$; $t_2 = -0.25$; $t_3 = -3.51$, and $R^2 = 0.9917$. Since there are 9 observations and 4 parameters there are $9 - 4 = 5$ degrees of freedom for error. The F value is

$$F = \frac{R^2}{1 - R^2} \left(\frac{5}{3} \right) = 119.1 \quad (7.21)$$

and this is significant at the 1% level since $f_{3,4,0.01} = 12.1$. Thus we can conclude that the overall fit is significant.

To test the assumption that there are two distinct lines we test

$$H_0 : \beta_3 = 0 \text{ against } H_0 : \beta_3 \neq 0. \tag{7.22}$$

Since $t_3 = -3.51$ and $t_{5,0.025} = 2.571$, H_0 can be rejected at the 5% level and it is quite reasonable to assume that the data is represented by two, rather than one line.

To check whether $x = 5$ is the abscissa of the point of intersection we test

$$H_0 : \beta_2 = 0 \text{ against } H_0 : \beta_2 \neq 0. \tag{7.23}$$

Since $t_2 = -0.25$ it is reasonable to conclude that $\beta_2 = 0$ and therefore that the true model is of the form

$$Y = \beta_0 + \beta_1 (x - 5) + \beta_3 (x - 5)_+ + \varepsilon. \tag{7.24}$$

We leave it as an exercise to refit the model under the assumption that the abscissa of the point of intersection is $x_0 = 5$.

Example 7.2 As a second example, consider the data set shown in Table 7.4 and plotted in Figure 7.5.

Table 7.4 Some data									
x	1	2	3	4	5	6	7	8	9
y	1.8	4.3	5.6	8.2	9.1	10.7	11.5	12.2	14.0

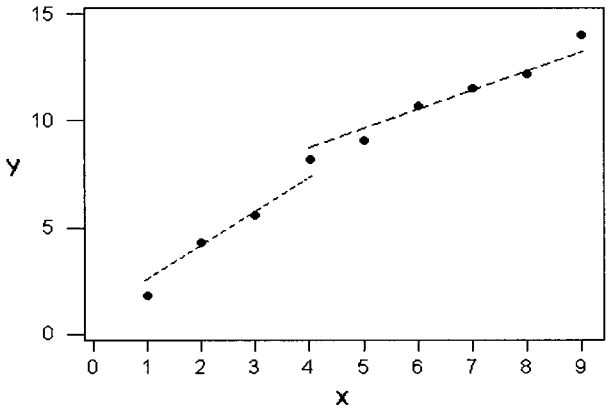


Figure 7.5: Scatter plot of x and y

Here it appears that the first 4 points lie on one line while the last five points appear to lie on another. Thus we assume a model of the form

$$Y = \beta_0 + \beta_1 (x - 4) + \beta_2 (x - 4)^0 + \beta_3 (x - 4)_+ + \varepsilon. \tag{7.25}$$

The data for this model are shown in Table 7.5.

Table 7.5 Data for Example 7.2

y	x_0	x_1	x_2	x_3
1.8	1	-3	0	0
4.3	1	-2	0	0
5.6	1	-1	0	0
8.2	1	0	0	0
9.1	1	1	1	1
10.7	1	2	1	2
11.5	1	3	1	3
12.2	1	4	1	4
14.0	1	5	1	5

Fitting this model by least squares gives:

$$\begin{aligned} \hat{\beta}_0 &= 8.05, & \hat{\beta}_1 &= 2.05, & \hat{\beta}_2 &= 0.06, & \hat{\beta}_3 &= -0.92, \\ t_0 &= 25.98, & t_1 &= 12.38, & t_2 &= 0.12, & t_3 &= -4.53, \\ R^2 &= 0.995, & \text{and } F &= 311.94. \end{aligned}$$

Thus the overall fit is significant at the 1% level and we conclude that $\hat{\beta}_2 \neq 0$ at the 1% level of significance. From our previous discussion, the estimated slope of the first line is $\beta_1 = 2.05$ while that of the second line is $\hat{\beta}_1 + \hat{\beta}_3 = 2.05 - 0.92 = 1.13$. Thus the estimated equation of the first line is

$$\hat{y}_1 = -0.15 + 2.05x,$$

while that of the second line is

$$\hat{y}_2 = 3.59 + 1.13x.$$

The abscissa of the point of intersection of the two lines is obtained by setting $\hat{y}_1 = \hat{y}_2$ and this gives

$$3.59 + 1.13x = -0.15 + 2.05x \quad \text{or} \quad 0.92x = 3.74$$

which yields $x = 4.065$.

The approach we have described for fitting a piecewise linear curve can be extended to fit more complicated piecewise polynomial models. For example, to fit data described by three lines over the intervals $(-\infty, x_0)$, $[x_0, x_1)$, (x_1, ∞) we can use the linear model

$$\begin{aligned} Y &= \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)_+^0 + \beta_3(x - x_0)_+ \\ &\quad + \beta_4(x - x_1)_+^0 + \beta_5(x - x_1)_+ + \varepsilon. \end{aligned} \quad (7.26)$$

Equality of slopes may be checked by testing

$$H_0 : \beta_3 = \beta_5 = 0 \quad (7.27)$$

against

$$H_1 : \beta_3 \neq 0 \quad \text{or} \quad \beta_5 \neq 0 \quad (7.28)$$

and this may be done by using the appropriate F test as described in Chapter 5.

A piecewise quadratic model may be obtained by fitting

$$\begin{aligned} Y = & \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \beta_3(x - x_0)_+^0 \\ & + \beta_4(x - x_1)_+ + \beta_5(x - x_1)_+^2 + \varepsilon. \end{aligned} \quad (7.29)$$

where

$$(x - x_0)_+^2 = \begin{cases} 0, & x \leq x_0, \\ (x - x_0)^2, & x > x_0. \end{cases} \quad (7.30)$$

In this form continuity of y at $x = x_0$ can be checked by testing

$$H_0 : \beta_3 = 0 \text{ against } H_1 : \beta_3 \neq 0 \quad (7.31)$$

while differentiability at $x = x_0$ may be checked by testing

$$H_0 : \beta_4 = 0 \text{ against } H_1 : \beta_4 \neq 0. \quad (7.32)$$

We leave the verification of these as exercises for the reader to check.

More complicated piecewise polynomial models may be developed along the lines we have indicated. For more details, we refer the reader to Refs. [87, 117].

One last comment should be made concerning the fitting of piecewise polynomial models. For instance in both of the examples given in this section, we assumed that we knew which points lay on which lines. In general, this will probably not be the case, so that we will have to estimate x_0 as well. One possible approach is to check every possible division of points into two lines. For example, in Example 7.2 we can take $x_0 = 1, 2, \dots, 9$ and the fit the model given by Eq. (7.25) for each possible choice of x_0 and then choose that value of x_0 which gives the largest R^2 (equivalently smallest SSE). For further details see Ref. [87].

Another possibility is to consider x_0 to be an unknown parameter and then one can estimate x_0 along with the coefficients. This is a nonlinear regression problem - but is complicated by the fact that the mean function is not differentiable at x_0 so traditional numerical methods for minimization based on calculus techniques are generally not applicable.

7.2.3 Multivariate Polynomial Models

If we are consider a polynomial regression model with two or more regressor variables, the approach will be a straightforward extension from the method of fitting polynomial models in one variable. Suppose that the postulated model is a second-order polynomial model in two variables. Then the model would be

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 + \varepsilon \quad (7.33)$$

where β_1 and β_2 are the two linear effects, β_{11} and β_{12} are the quadratic effects of x_1 and x_2 respectively, and β_{12} indicates the parameter of an interaction effect. Then the *regression function* (or *response function*)

$$E(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 \quad (7.34)$$

is called a *response surface*. This type of modeling is also called *response surface methodology*, which is widely applicable to model output processes in industrial or engineering areas for finding the operation conditions to optimize a response. For more details and examples, see [10, 88].

7.3 Radial Basis Functions

Although most regression models used in practice are parametric, i.e., the parameter vector has a direct physical interpretation, there are many situations where they do not. Such models are often referred to as *nonparametric*. Typical examples are polynomial and spline models. Even though splines are quite useful for fitting global data, they can be clumsy to use for higher dimensional data. Similarly, polynomials can cause difficulties if the range of the data is large and the dimensionality is high. To remedy some of these problems, mathematicians have been investigating a new class of functions, *radial basis functions* (rbfs). Although much of this work has been devoted to interpolating non-noisy data, increasingly they have been used to fit noisy data sets by statistical, often least squares methods.

Typical applications, include approximation of spatial data as occurs in mining studies, particularly, in the well-known statistical technique of *kriging* [23], environmental data, medical data and neural network modeling.

To further elaborate on these functions we begin with a definition.

Definition 7.1 Let $\phi : [0, \infty) \rightarrow \mathbb{R}$ be a continuous real-valued function. Let $\|\cdot\|$ denote the Euclidean norm on \mathbb{R}^n and let $\{\mathbf{P}_j\}_{j=1}^N$ be a set of N distinct points in \mathbb{R}^n . A radial basis function is a function of the form

$$f(\mathbf{P}) = \sum_{j=1}^N \beta_j \phi(\|\mathbf{P} - \mathbf{P}_j\|), \quad \mathbf{P} \in \mathbb{R}^n \quad (7.35)$$

where $\{\beta_j\}_{j=1}^N$ are unknown coefficients. These coefficients are generally determined from a given set of values of f at a possibly distinct set of points $\{\mathbf{Q}_k\}_{k=1}^M$, $M \geq N$.

More generally, one often adds a polynomial term p_m of degree m to (7.35) so that the most general rbf is given by

$$f(\mathbf{P}) = \sum_{j=1}^N \beta_j \phi(\|\mathbf{P} - \mathbf{P}_j\|) + p_m. \quad (7.36)$$

Then, the coefficients of p_m have to be determined simultaneously with $\{\beta_j\}_{j=1}^N$.

Although by definition, there is an infinite number of possibilities for ϕ , over the years a relatively small number of ϕ 's have emerged as being particularly useful.

7.3.1 Types of Radial Basis Functions

Let $r = \|\mathbf{P}\|$, then typical radial basis functions are of the form;

(i) $\phi(r) = \exp(-cr^2)$, $c > 0$, $\mathbf{P} \in \mathbb{R}^n$.

These functions are referred to as *Gaussian radial basis functions*. It is known that Gaussian rbfs have optimal convergence properties and have found widespread use in neural network modeling [95, 100]. The parameter c is usually called the *shape* or *variance parameter*. Generally its value is unknown and its proper choice can have a substantial effect on the quality of the fit.

(ii) $\phi(r) = \sqrt{r + c^2}$, $\mathbf{P} \in \mathbb{R}^n$.

Here $\phi(r)$ is called a *multiquadric* and c is the shape parameter. Again multiquadrics have optimal convergence properties, with the shape parameter having a marked effect on the convergence rate and its proper choice has been a topic of continuing interest. Multiquadrics have found considerable use in fitting spatial data [36, 78] and in solving partial differential equations [43]. Interestingly they were first discovered by Hardy [50] in the course of fitting geophysical data.

(iii) $\phi(r) = \begin{cases} r^2 \log r, & \mathbf{P} \in \mathbb{R}^2, \\ r, & \mathbf{P} \in \mathbb{R}^3. \end{cases}$

These functions are referred to as *thin-plate splines* (TPS) and are probably the most widely used rbfs. They are generally considered to be the first rbfs to be actively investigated and have found widespread use in fitting spatial data (see in particular, Whaba [117]), fitting technical data and in the solution of partial differential equations [42]. Their importance derives from the fact that they are optimal interpolants and are the natural generalization of one dimensional cubic splines [117].

In contrast to Gaussians and multiquadrics, which do not require the addition of polynomial terms, TPS require the condition of a first degree polynomial

$$p_1(\mathbf{P}) = a + bx + cy, \mathbf{P} = (x, y) \text{ in } \mathbb{R}^2 \quad (7.37)$$

and

$$p_1(\mathbf{P}) = a + bx + cy + dz, \mathbf{P} = (x, y, z) \text{ in } \mathbb{R}^3. \quad (7.38)$$

In this case, in order to properly fit the data it is necessary to impose the constraints

$$\sum_{j=1}^N \beta_j = \sum_{j=1}^N \beta_j x_j = \sum_{j=1}^N \beta_j y_j = 0, (x_j, y_j) \in \mathbb{R}^2, \quad (7.39)$$

$$\sum_{j=1}^N \beta_j = \sum_{j=1}^N \beta_j x_j = \sum_{j=1}^N \beta_j y_j = \sum_{j=1}^N \beta_j z_j = 0, (x_j, y_j, z_j) \in \mathbb{R}^3. \quad (7.40)$$

(iv) $\phi(r) = \begin{cases} r^{2n} \log r, & \mathbf{P} \in \mathbb{R}^2, n \geq 2 \\ r^{2n-1}, & \mathbf{P} \in \mathbb{R}^3, n \geq 2. \end{cases}$

These rbfs are called *higher-order polyharmonic* or *Duchon splines*. These rbfs are a direct generalization of thin plate splines (TPS) and generally their approximation power increases with n , but this is counteracted by the need to add a polynomial of degree n which causes increased ill-conditioning as n increases. As for TPS, there rbfs can be obtained from a least squares principle [117]. Although they have been known for almost 30 years, they have yet to find widespread use in practice - most analysts generally settle for the simpler TPS [116].

- (v) Over the years, a persistent criticism of rbfs has been the fact that the well-known rbfs have global support and so matrices associated with their approximation are not sparse. For many years, the holy grail of rbf theory was to find a class of rbfs with compact support and whose related interpolation matrices are invertible for arbitrary data sets.

In the mid nineteen-nineties this problem was resolved by Wu [124] and improved upon by Wendland in [119]. Without going into details all of these functions have the form

$$\phi(r) = \begin{cases} (1-r)_+^n p(r), & 0 \leq r \leq 1, \\ 0, & r > 1, \end{cases} \quad (7.41)$$

where

$$(1-r)_+^n = \begin{cases} (1-r)^n, & 0 \leq r \leq 1, \\ 0, & r > 1, \end{cases} \quad (7.42)$$

and $p(r)$ is a polynomial whose degree depends on the dimension of the data space.

If one scales r by, $r \rightarrow r/a$, then we obtain rbfs supported on $0 \leq r \leq a$. From (7.41), it follows that

$$\phi_a(r) = \begin{cases} (1-r/a)_+^n p(r/a), & 0 \leq r \leq a, \\ 0, & r > a. \end{cases} \quad (7.43)$$

As for Gaussians and multiquadrics, the scale parameter a has a substantial effect on the accuracy of the approximation by $\phi_a(r)$. As a increases, the interpolation matrices become less sparse, while their approximation properties improve. Finding the ‘optimal’ value of a to balance these competing effects is a difficult and not a totally solved problem. Some methods for doing this can be found in [43].

We next turn to methods for determining the various parameters in (7.35) and (7.36).

7.3.2 Fitting Methods for RBFs

At present, there are three major methods for determining the unknown parameters in rbf approximations: *interpolation* [96], *smoothing interpolation* [117] and *linear and nonlinear least squares* [100]. Generally, one uses interpolation in the non-statistical context (but not always, since the well-known statistical technique of kriging is known to be equivalent to rbf interpolations). The latter two techniques may be viewed as generalizations of interpolation. Hence, we begin with a brief discussion of interpolation and to simplify matters, we restrict ourselves to rbfs where the polynomial $p_m = 0$. These include Gaussians, multiquadrics and Wendland’s compactly supported rbfs and rbfs such as the inverse multiquadrics $\phi(r) = (r^2 + c^2)^{-1/2}$.

Interpolation

For interpolation we assume that $\mathbf{Q}_k = \mathbf{P}_k$, $1 \leq k \leq N$ and assume that we know $y_k = f(\mathbf{P}_k)$. (Sometimes f is known, but usually in the statistical context it is not.) In this case we equate the right hand side of (7.35) to y_k , $1 \leq k \leq N$. This gives

$$\sum_{j=1}^N \beta_j \phi(\|\mathbf{P}_k - \mathbf{P}_j\|) = y_k, \quad 1 \leq k \leq N. \quad (7.44)$$

Letting $\beta = (\beta_1, \beta_2, \dots, \beta_N)^T$

$$\Phi = [\phi(\|\mathbf{P}_k - \mathbf{P}_j\|)], 1 \leq j \leq N, 1 \leq k \leq N \quad (7.45)$$

and $\mathbf{y} = (y_1, y_2, \dots, y_N)^T$, then (7.44) can be written in matrix-vector form

$$\Phi\beta = \mathbf{y}. \quad (7.46)$$

For Gaussians and the compactly supported rbfs, Φ is positive definite [119, 95], hence, invertible. For multiquadrics Φ is invertible, but not positive definite while for inverse multiquadrics Φ is positive definite. Hence,

$$\beta = \Phi^{-1}\mathbf{y} \quad (7.47)$$

so that β is uniquely determined.

Smoothed Interpolation

When the data are noisy, interpolation may not be appropriate. In this case, rather than choosing β to exactly satisfy (7.44), we attempt to smooth the fit by preventing the approximation from exactly passing through the data. Note first that interpolation is equivalent to minimizing the residual sum of squares.

$$L = \langle \mathbf{y} - \Phi\beta, \mathbf{y} - \Phi\beta \rangle \quad (7.48)$$

with respect to β .

Adding a penalty term of the form $\lambda \langle \beta, \beta \rangle = \lambda \|\beta\|^2$, $\lambda > 0$ to L , we now determine β by minimizing

$$L' = \langle \mathbf{y} - \Phi\beta, \mathbf{y} - \Phi\beta \rangle + \lambda \|\beta\|^2 \quad (7.49)$$

with respect to β . In this form this approach is equivalent to the *ridge regression* problem discussed in Section 9.5. As shown there, the problem of choosing the *ridge* or *smoothing parameter* λ is a non-trivial problem and much of the theory over the past thirty years has centered on that problem.

In Wahba's work, she focuses on the use of *cross-validation* which is briefly discussed in Section 9.5. Further details can be found in [117].

Least Squares Fitting

In this approach we assume that the data points $\{\mathbf{Q}_k\}_{k=1}^M$, $M > N$, are generally distinct from the centers $\{\mathbf{P}_j\}_{j=1}^N$. In this case the rbf model can be written in standard regression form

$$y_k = \sum_{j=1}^N \beta_j \phi(\|\mathbf{P}_j - \mathbf{Q}_k\|) + \varepsilon_k, 1 \leq k \leq M \quad (7.50)$$

where $\{\varepsilon_k\}_{k=1}^M$ are independent $N(0, \sigma^2)$ random variables. Using our previous notation (7.50) can be written in standard regression form as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (7.51)$$

where

$$\mathbf{X} = [\phi(\|\mathbf{P}_j - \mathbf{Q}_k\|)], \quad 1 \leq j \leq N, \quad 1 \leq k \leq M. \quad (7.52)$$

Hence β can be determined by minimizing the residual sum of squares

$$L = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle. \quad (7.53)$$

Here there are number of distinct possibilities. If $\{\mathbf{P}_j\}_{j=1}^N$ and the shape parameters are also known, then minimizing L is a standard regression problem and $\hat{\beta}$ is given by (5.22a), i.e.,

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (7.54)$$

where $\mathbf{X} = [\phi(\|\mathbf{P}_j - \mathbf{Q}_k\|)], \quad 1 \leq j \leq N, \quad 1 \leq k \leq M$, and an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\langle \mathbf{y} - \mathbf{X}\hat{\beta}, \mathbf{y} - \mathbf{X}\hat{\beta} \rangle}{N - M}. \quad (7.55)$$

One can then use all of the machinery of linear regression theory to evaluate the adequacy of the model.

In the second case, the centers $\{\mathbf{P}_j\}_{j=1}^N$ are not known and must be determined from the data. Generally, the shape parameters will be unknown as well. As before, we minimize

$$L = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle \quad (7.56)$$

with respect to β . However the centers and shape parameters appear nonlinearly in L and we minimize L with respect to $\{\mathbf{P}_j\}_{j=1}^N$ and the shape parameters c simultaneously. This is then a *nonlinear regression* problem and can be solved by solving the non-linear normal equations:

$$\begin{cases} \partial L / \partial \beta_j = 0, & 1 \leq j \leq N, \\ \partial L / \partial \mathbf{P}_j = 0, & 1 \leq j \leq N, \\ \partial L / \partial c = 0. \end{cases} \quad (7.57)$$

Unfortunately, these equations generally do not have an analytic solution as in the linear case. In this case (7.57) have to be solved by some numerical iterative method, such as the Gauss-Newton method. Further details can be found in [27].

7.4 Dummy Variables

As we have already pointed out in Chapter 5, one of the important properties of the GLM is its ability to incorporate both *qualitative* as well as *quantitative variables* in the model. In many areas such as medicine and social sciences, there are often many more qualitative factors than quantitative ones and the ability to account for the effects of these factors is one of the great attractions of using the GLM. In addition to the effect of gender discussed in Example 5.15 qualitative factors of all sorts occur in scientific studies. In economics, seasonality factors are important and sociological studies often require one to account for geographical differences. In [64] the authors considered the effects of various factors affecting the survival of patients following admission of patients to a hospital intensive care unit. Among qualitative factors considered were: sex, race, presence of cancer, history of renal failure, previous admission, ph from blood gases and many others. Of 19 independent variables only three were quantitative.

In this section we expand on our work in Chapter 5 showing how to incorporate multiple qualitative factors, qualitative factors at more than two levels and interactions into the GLM. In addition, we give a brief comparison of the use of dummy variables and the more traditional analysis of variance. We begin by recapitulating some ideas from Chapter 5.

Consider the following hypothetical salary data for eight professors at “good ole” B.S.U.

Table 7.6 Hypothetical Salary data

Rank	No.	Salary (in thousands of \$)
Associate Professor	1	22.5
	2	33.5
	3	25.0
	4	27.0
Full Professor	5	39.0
	6	40.0
	7	37.0
	8	36.0

We would like to determine whether there is a difference in the mean salaries of full and associate professors. The standard approach to such a problem is to use the t -test for testing the difference between two means. An alternative (but equivalent) approach which has significant generalizations is to make a linear model for the professors’ salaries by using dummy variables and then utilizing a standard t -test for a regression coefficient.

Now the variable of interest here is the rank of a professor, a qualitative variable. We quantify this variable in the following way: Let

$$x = \begin{cases} 0, & \text{if the professor is an associate,} \\ 1, & \text{if the professor is full.} \end{cases}$$

If Y denotes the salary of an arbitrary professor then (assuming that rank is the only explanatory variable) we claim that

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (7.58)$$

where $\beta_0 = E(Y_{\text{associate}})$ and $\beta_0 + \beta_1 = E(Y_{\text{full}})$ so that

$$\beta_1 = E(Y_{\text{full}}) - E(Y_{\text{associate}}) \quad (7.59)$$

represents the difference in the mean salaries of full and associate professors.

To verify this, observe that if we take $x = 0$, then $E(Y) = \beta_0$, and if $x = 1$ then $E(Y) = \beta_0 + \beta_1$. If the errors are normal with constant variance, then all inferences concerning the professors’ salaries can be made using the linear model. For example, differences in the mean salaries can be checked by testing

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0. \quad (7.60)$$

This test can be shown to be equivalent to the usual t -test for the difference of two normal means. The parameter estimates (as follows from (3.18) and (3.19)) have their expected values; i.e.,

$$\hat{\beta}_0 = \frac{1}{4} \sum_{i=1}^4 x_{i(\text{associate})} = 27.0 \quad (7.61)$$

and

$$\hat{\beta}_0 + \hat{\beta}_1 = 38.0, \quad (7.62)$$

which is the average value of the full professor's salaries. We also find that

$$t_0 = 15.123; \quad t_1 = 4.3566 \quad (7.63)$$

where t_1 has the same value as the classical t -statistic for testing the difference of two normal means. From this it follows that the difference in mean salaries is significant at the 5% level.

We now consider extending this model for the purpose of testing whether there is a difference among the mean salaries for full, associate and assistant professors. Again we try to make a salary model which depends only on the rank of a professor. We do this by introducing two dummy variables as follows:

$$\begin{aligned} x_1 &= \begin{cases} 1, & \text{if professor is an associate,} \\ 0, & \text{otherwise,} \end{cases} \\ x_2 &= \begin{cases} 1, & \text{if professor is an assistant,} \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

The salary model is now of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (7.64)$$

where

$$\begin{aligned} E(Y_{full}) &= \beta_0, \\ E(Y_{associate}) &= \beta_0 + \beta_1, \\ E(Y_{assistant}) &= \beta_0 + \beta_2. \end{aligned}$$

To see this, let $x_1 = x_2 = 0$, then $E(Y) = \beta_0$. But $x_1 = x_2 = 0$ if and only if a professor is a full professor. Similarly, if $x_1 = 1$ and $x_2 = 0$, this indicates an associate professor and then $E(Y) = \beta_0 + \beta_1$, while for an assistant professor $x_1 = 0$, $x_2 = 1$ and $E(Y) = \beta_0 + \beta_2$. Thus β_1 and β_2 measure the difference between full and associate salaries and full and assistant salaries respectively. The test for overall differences in salaries is

$$H_0 : \beta_1 = \beta_2 = 0 \quad (7.65)$$

against

$$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0. \quad (7.66)$$

This may be done using the overall F test for the significance of the regression. This test is equivalent to the classical one-way analysis variance test for testing the difference between three means. The model given in (7.64) is then a particular case of the one-way ANOVA model.

The general case of a qualitative factor at k levels may be treated by generalizing the discussion given above. We do this by introducing $k - 1$ (0-1)-valued dummy variables in the following way: Let

$$x_j = \begin{cases} 1, & \text{if the observation is in the } j\text{-th category,} \\ 0, & \text{otherwise,} \end{cases} \quad (7.67)$$

$j = 1, 2, \dots, k - 1$. The model for Y is then given by

$$Y = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_j + \varepsilon, \tag{7.68}$$

where the expected value of observations in the k -th category is β_0 and those in the j -th category, $1 \leq j \leq k - 1$, have expected values $\beta_0 + \beta_j$. (The k -th category is often referred to as the *excluded category*.) To test whether there is a difference between the means of the k categories we test

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{k-1} = 0 \tag{7.69}$$

against

$$H_1 : \text{at least one } \beta_j \neq 0, \ 1 \leq j \leq k - 1. \tag{7.70}$$

This is nothing other than the F test for the overall significance of the regression.

Example 7.3 We now consider an additional 4 assistant professors whose salaries are listed in the Table 7.7.

Table 7.7 Salary data of Assistant Professor		
Rank	No.	Salary (in thousands of \$)
Assistant Professor	9	21.0
	10	23.0
	11	30.0
	12	25.0

If we introduce dummy variables as before, the data matrix for the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{7.71}$$

is

Table 7.8 Salary data using Dummy Variables				
Rank	y	x_0	x_1	x_2
Full Professor	39.0	1	0	0
	40.0	1	0	0
	37.0	1	0	0
	36.0	1	0	0
Associate Professor	22.5	1	1	0
	33.5	1	1	0
	25.0	1	1	0
	27.0	1	1	0
Assistant Professor	21.0	1	0	1
	23.0	1	0	1
	30.0	1	0	1
	25.0	1	0	1

Fitting this model by least squares gives the results:

$$\begin{aligned} \hat{\beta}_0 &= 38.0, & \hat{\beta}_1 &= -11.0, & \hat{\beta}_2 &= -13.0, \\ t_0 &= 20.710, & t_1 &= -4.238, & t_2 &= -5.105, \\ R^2 &= 0.7683, & \text{and } F &= 14.925. \end{aligned}$$

From this we see that the overall regression is significant at the 1% level so we reject the hypothesis of no differences between the mean salaries of professors of different ranks.

We also note (as may be checked) that $\hat{\beta}_0 = 38.0$ is the average salary of the 4 full professors, $\hat{\beta}_0 + \hat{\beta}_1 = 27.0$ is the average salary of the 4 associate professors and $\hat{\beta}_0 + \hat{\beta}_2$ is the average salary of the 4 assistant professors.

Example 7.4 As a further example of the use of dummy variables we consider the results of an agricultural experiment. Such data are traditionally analyzed using ANOVA methods.

Five varieties of peas were planted, each on 4 different plots. The yields in bushels per acre are shown in Table 7.9:

Table 7.9 Pea Harvest Data					
Yield	Variety				
	A	B	C	D	E
#1	26.2	29.2	29.1	21.3	20.1
#2	24.3	28.1	30.8	22.4	19.3
#3	21.8	27.3	33.9	24.3	19.9
#4	28.1	31.2	32.8	21.8	22.1

To test whether there are differences in the yields of the five varieties we consider a linear model for the yield per acre as

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \varepsilon \tag{7.72}$$

where the x_i 's are dummy variables for varieties B-E with A being taken as the excluded category.

To fit this model we use the data given in Table 7.10.

Table 7.10 Pea Data using Dummy Variables for Variety						
Variety	y	x_0	x_1	x_2	x_3	x_4
A	26.2	1	0	0	0	0
	24.3	1	0	0	0	0
	21.8	1	0	0	0	0
	28.1	1	0	0	0	0
B	29.2	1	1	0	0	0
	28.1	1	1	0	0	0
	27.3	1	1	0	0	0
	31.2	1	1	0	0	0
C	29.1	1	0	1	0	0
	30.8	1	0	1	0	0
	23.9	1	0	1	0	0
	32.8	1	0	1	0	0
D	21.3	1	0	0	1	0
	22.4	1	0	0	1	0
	24.3	1	0	0	1	0
	21.8	1	0	0	1	0
E	20.1	1	0	0	0	1
	19.3	1	0	0	0	1
	19.9	1	0	0	0	1
	22.1	1	0	0	0	1

This model was fit by least squares with the following results;

$$\begin{aligned}\hat{\beta}_0 &= 25.1, & \hat{\beta}_1 &= 3.85, & \hat{\beta}_2 &= 6.55, & \hat{\beta}_3 &= -2.65, & \hat{\beta}_4 &= -4.75, \\ t_0 &= 26.6, & t_1 &= 2.88, & t_2 &= 4.90, & t_3 &= -1.98, & t_4 &= -3.56, \\ R^2 &= 0.8647, & \text{and } F &= 23.97.\end{aligned}$$

Here we can check that $\hat{\beta}_0$ is the mean value of the yields of variety A. Similarly, $\hat{\beta}_0 + \hat{\beta}_1 = 28.95$ is the mean yield for B, $\hat{\beta}_0 + \hat{\beta}_2 = 31.65$ is the mean yield for C, $\hat{\beta}_0 + \hat{\beta}_3 = 22.45$ is the mean yield for D while $\hat{\beta}_0 + \hat{\beta}_4 = 20.35$ is the mean yield for E.

Since there are 4 degrees of freedom for regression and 15 degrees of freedom for error, we find that $f_{4,15,0.01} = 4.89$ so that we can reject H_0 at the 1% level and conclude that differences in yields are highly significant.

Of course, we may consider models with two or more qualitative factors. If, for example we have two variables, one with j levels and the other with k levels, the resultant model is usually called a *two-way analysis of variance*.

For m qualitative variables, the model is *m-way analysis of variance* with $m = 2$ being perhaps the most common case. As an example suppose we consider the effect of gender on professors's salaries. It turns out that the sex distribution is:

Table 7.11 Rank and Gender of Professor

Rank of Professor	No.	Sex
Full	1	F
	2	F
	3	F
	4	M
Associate	5	F
	6	F
	7	F
	8	F
Assistant	9	F
	10	M
	11	M
	12	M

To account for this additional variable we introduce another dummy variable

$$x_3 = \begin{cases} 0, & \text{if male,} \\ 1, & \text{if female.} \end{cases}$$

The model is now of the form

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{\text{rank}} + \underbrace{\beta_3 x_3}_{\text{sex}} + \varepsilon \quad (7.73)$$

where

- β_0 = mean salary of a male full professor,
- $\beta_0 + \beta_1$ = mean salary of a male associate professor,
- $\beta_0 + \beta_2$ = mean salary of a male assistant professor,
- $\beta_0 + \beta_3$ = mean salary of a female full professor,
- $\beta_0 + \beta_1 + \beta_3$ = mean salary of a female associate professor,
- $\beta_0 + \beta_2 + \beta_3$ = mean salary of a female assistant professor.

Thus, (x_1, x_2) measures the effect of rank and x_3 measures the effect of sex. The effect of sex can be determined by testing $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$ using the usual t test. To do this, we fitted the model (7.73) using the data in Table 7.12.

Table 7.12 Salary Data using Dummy Variables

Rank	y	x_0	x_1	x_2	x_3
Full	39.0	1	0	0	1
	40.0	1	0	0	1
	37.0	1	0	0	1
	36.0	1	0	0	0
Associate	22.5	1	1	0	1
	33.5	1	1	0	1
	25.0	1	1	0	1
	27.0	1	1	0	1
Assistant	21.0	1	0	1	1
	23.0	1	0	1	0
	30.0	1	0	1	0
	25.0	1	0	1	0

The results were;

$$\begin{aligned}\hat{\beta}_0 &= 38.88, & \hat{\beta}_1 &= -10.71, & \hat{\beta}_2 &= -13.83, & \hat{\beta}_3 &= -1.17, \\ t_0 &= 12.74, & t_1 &= -3.77, & t_2 &= -4.39, & t_3 &= -0.37, \\ R^2 &= 0.7722, \text{ and } F = 9.04.\end{aligned}$$

From this we see that the overall regression is significant at the 1% level but the small value of t_3 indicates that sex is not a significant predictor. (This is not surprising since the example was made up by randomly assigning the sex of the professor by tossing a coin.)

Of course a model may contain quantitative as well as qualitative factors, faculty salaries may depend on age, number of publications, length of service and many others as well. If these factors contribute linearly then they will be added to the model, just as the qualitative factors. So for a linear model for professors' age x_4 and number of publications x_5 , the model would be

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon. \quad (7.74)$$

Example 7.5 Let's consider the jewelry sales data in Example 3.21. To adjust for seasonal trends, one of the common methods is to take a moving average (over a whole year) of the time series. Here we introduce dummy variables as a tool to adjust for the seasonality in the data. Now consider the model with three dummy variables, q_2, q_3 and q_4 , as follows:

$$Y = \beta_0 + \beta_1 x_1 + (\beta_2 q_2 + \beta_3 q_3 + \beta_4 q_4) + \varepsilon \quad (7.75)$$

where q_i ($i = 2, 3, 4$) indicates the i -th quarter effect in sales. In fact, q_i measures the shift from a first quarter base. Then, the model (7.75) becomes

- (i) $Y = \beta_0 + \beta_1 x_1 + \varepsilon$; for the effect of the first quarter,
- (ii) $Y = \beta_0 + \beta_1 x_1 + \beta_2 q_2 + \varepsilon$; for the effect of the second quarter,
- (iii) $Y = \beta_0 + \beta_1 x_1 + \beta_3 q_3 + \varepsilon$; for the effect of the third quarter,
- (iv) $Y = \beta_0 + \beta_1 x_1 + \beta_4 q_4 + \varepsilon$; for the effect of the fourth quarter.

The data for estimating β is given in Table 7.13.

Table 7.13 Jewelry Sales data using Dummy Variables					
Year	Quarter (x_1)	Sales Y (in \$100,000)	q_2	q_3	q_4
1957	1	36	0	0	0
	2	44	1	0	0
	3	45	0	1	0
	4	106	0	0	1
1958	1	38	0	0	0
	2	46	1	0	0
	3	47	0	1	0
	4	112	0	0	1
1959	1	42	0	0	0
	2	49	1	0	0
	3	48	0	1	0
	4	118	0	0	1
1960	1	42	0	0	0
	2	50	1	0	0
	3	51	0	1	0
	4	118	0	0	1

This model was fit by least squares with the following results;

$$\begin{aligned} \hat{\beta}_0 &= 34.95, & \hat{\beta}_1 &= 0.65, & \hat{\beta}_2 &= 7.10, & \hat{\beta}_3 &= 6.95, & \hat{\beta}_4 &= 72.05, \\ t_0 &= 32.12, & t_1 &= 6.79, & t_2 &= 5.84, & t_3 &= 5.67, & t_4 &= 57.86, \\ \bar{R}^2 &= 0.998, & \text{and } F &= 1230.41. \end{aligned}$$

From the values of the t -statistics and the coefficient of determination in the results, and the fact that $F = 1230.41 > f_{4,11,0.01}$ the overall regression is significant at the 1% level. Hence, we conclude that the dummy variables adequately explain the seasonality of the quarterly sales in the data. For example, we are able to know that fourth quarter dominates the seasonality in the sales because the average seasonal shift for the fourth quarter is, $\hat{\beta}_4 = 72.05$.

The advantage of using dummy variables is that both seasonal shifts and the relationship of sales (Y) to time (x_1) are estimated simultaneously in the same regression model. We also note that the slope $\hat{\beta}_1$ is now an unbiased estimate of β_1 .

In Figure 7.6, the residuals from the model (7.75) are now rather evenly scattered from zero in which the two groups represent the first three quarters (low-sales season) and fourth quarter (high-sales season) respectively. The residual plot seems to show that the errors are uncorrelated which is probably accounted for by the adjustment for seasonality in the model.

7.4.1 Further Comments on Dummy Variable

Although we have discussed dummy variables using 0-1 coding it is possible to use other coding as well. For example, qualitative variables, with two levels are often coded 1-2 rather than 0-1. Generally 0-1 coding is preferred because the regression coefficients are usually easier to interpret and the effect of multicollinearity is reduced. For a factor at

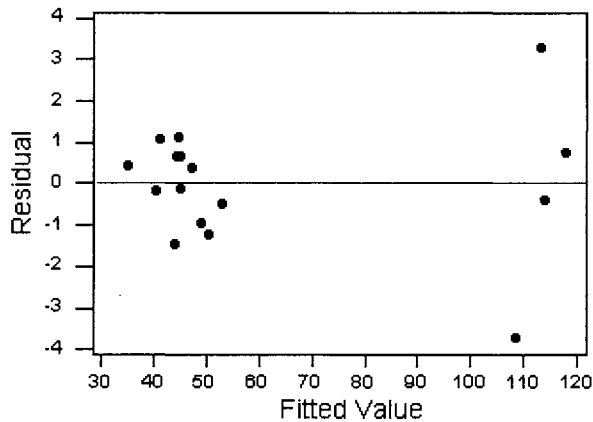


Figure 7.6: Residuals versus fitted values for jewelry sales data

three levels 0-1 coding can be displayed by a 3×2 table

	x_1	x_2
level 1	1	0
level 2	0	1
level 3	0	0

(7.76)

Note that the columns are linearly independent vectors so that arbitrary codings may be used as long as the possible values are linearly independent. For example,

	x_1	x_2
level 1	1	0
level 2	-1	1
level 3	0	0

(7.77)

is permissible while

	x_1	x_2
level 1	1	-1
level 2	1	-1
level 3	0	0

(7.78)

is not.

This is easily extended to a factor at k levels - the coding used must be given by $k - 1$ linearly independent vectors. In addition, one must be careful not to use too many dummies since that can inadvertently cause the model to have less than full rank. For example if we have a factor at 2 levels and use two dummies

	x_1	x_2
level 1	1	0
level 2	0	1

(7.79)

one can see that $x_1 + x_2 = 1$ so the columns of the design matrix are linearly dependent. An example of this can be found in [27].

If one wishes to do ANOVA using a standard regression model, then one needs to make sure that dummy variables are introduced which make the design matrix have full rank. On the other hand, traditional ANOVA models use linear models with overspecified variables resulting (by design) in design matrices which are less than full rank. In this case the standard form of the GLM cannot be used and additional constraints need to be introduced to allow the parameters to be estimated. This usually results in a constrained least squares problem. From this point of view, traditional ANOVA is mathematically more complicated than regression analysis.

Example 7.6 As an example of the above remarks we re-examine the professors salary data in Table 7.6. Let Y_{jk} be the salary of the k -th professor in the j -th category. Here, $j = 1$ for an associate and $j = 2$ for a full and $1 \leq k \leq 4$. In the traditional ANOVA approach we would use a model of the form

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}, \quad j = 1, 2, \quad k = 1, 2, 3, 4. \quad (7.80)$$

From Table 7.6 we get

$$\begin{aligned} 22.5 &= Y_{11} = \mu + \alpha_1 + \varepsilon_{11}, \\ 33.5 &= Y_{12} = \mu + \alpha_1 + \varepsilon_{12}, \\ 25.0 &= Y_{13} = \mu + \alpha_1 + \varepsilon_{13}, \\ 27.0 &= Y_{14} = \mu + \alpha_1 + \varepsilon_{14}, \\ 39.0 &= Y_{21} = \mu + \alpha_2 + \varepsilon_{21}, \\ 40.0 &= Y_{22} = \mu + \alpha_2 + \varepsilon_{22}, \\ 37.0 &= Y_{23} = \mu + \alpha_2 + \varepsilon_{23}, \\ 36.0 &= Y_{24} = \mu + \alpha_2 + \varepsilon_{24}. \end{aligned} \quad (7.81)$$

Letting $\mathbf{Y} = (Y_{11}, Y_{12}, \dots, Y_{24})^T$, $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2)^T$ and

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad (7.82)$$

the data can be represented as a linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (7.83)$$

However, notice that \mathbf{X} has rank two because the sum of the last two columns equals the first. Hence, $\mathbf{X}^T\mathbf{X}$ is not invertible so the coefficients $(\mu, \alpha_1, \alpha_2)$ cannot be obtained as before. In effect, the usual ANOVA model is over parameterized, having three rather than two parameters necessary to represent the differences in salary. The regression approach gives the correct number of variables for a nonsingular model.

To obtain an *estimable model* from (7.83) we need to impose a constraint on the parameters. Typically one uses $\alpha_1 + \alpha_2 = 0$. In this case we can eliminate α_2 from the model which can now be represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (7.84)$$

where $\beta = (\mu, \alpha_1)^T$ and

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix} \quad (7.85)$$

which is a full rank model.

Further examples will be given in the Exercises.

7.5 Interactions

As for quantitative variables one can have *interactions* between qualitative and quantitative variables. As we indicated in the Example 5.5 an interaction between a qualitative variable with two levels and a quantitative variable can be used to model the possibility that the slope of the quantitative variable is different for each level of the quantitative variable. Geometrically, this can be represented by a model representing two non-parallel lines. For a qualitative factor at k levels an interaction with a single quantitative variable implies that the model represents k non-parallel lines.

For example, in modeling professor salaries, suppose salaries depend on age, but the rate of change of salary (slope) depends on the professor's rank. Then, if x_1, x_2 are the dummy variables used to represent rank and if x_4 denotes age, the interaction between rank and age would be modeled by adding terms of the form

$$\beta_6 x_1 x_4 + \beta_7 x_2 x_4 \quad (7.86)$$

to the model (7.64). In general, an interaction (more specifically a *two-way interaction*) between a quantitative variable x_1 and a qualitative variable with k levels can be represented by adding terms of the form

$$\gamma_1 x_1 x_2 + \gamma_2 x_1 x_3 + \cdots + \gamma_{k-1} x_1 x_k, \quad (7.87)$$

where $x_j, 1 \leq j \leq k-1$ are the dummy variables used to represent the levels of the qualitative variable. That is, we enter into the model all possible products of the quantitative and qualitative variables. If x_1 is a qualitative variable having two levels then (7.87) can be coded to represent an interaction between it and the second factor having k levels. For example, an interaction between sex and rank would be represented by terms of the form in (7.73)

$$\gamma_1 x_1 x_3 + \gamma_2 x_2 x_3. \quad (7.88)$$

More generally, an interaction between a qualitative factor at k levels and another at j levels would be represented by adding all possible products of the dummies taken two at a time. For example, an interaction between a qualitative factor at three levels represented by two dummies x_1 and x_2 and one with four levels represented by x_3, x_4 and x_5 would be represented by adding a term of the form

$$\gamma_1 x_1 x_3 + \gamma_2 x_1 x_4 + \gamma_3 x_1 x_5 + \gamma_4 x_2 x_3 + \gamma_5 x_2 x_4 + \gamma_6 x_2 x_5 \quad (7.89)$$

to the model. In general these will be $(j - 1)(k - 1)$ new variables added to the model.

As one can see, adding interactions very quickly increases the number of independent variables and hence, the complexity of the model. To make matters worse, one may wish to consider 3-way, 4-way and higher order interactions into the model to cover all possible interactions between all the variables. If one wishes to be very cautious about the possible interactions between variables one would enter all of them to begin with and test to see if any are significant. Unfortunately, this can cause substantial difficulties, in both interpretation and computation.

For even a small number of initial variables, including all interactions can easily produce a model with more variables than observations and hence the assumptions of the GLM fail. Even if that does not occur, the introduction of interactions can cause substantial multicollinearity in the model, similar to that in using high order polynomial models for quantitative variables.

This can make the model difficult to interpret and may produce strange anomalies where t or F tests can show significant interaction effects but non significant direct effects. Moreover, some, but not all of the coefficients in an interaction could be significant but not others.

As a partial remedy for the multicollinearity one can use centered variables or partial orthogonalization as indicated in [27, 87]. To illustrate some of those possibilities we consider data given in [99].

Example 7.7 (Pulse Data [99]) In an experiment 92 students measured their pulse rate, then each student was asked to flip a coin. If the coin came up heads, then they were asked to run in place for one minute. Then everyone was asked to measure their pulse rates again. This second pulse rate was recorded and various other factors were recorded. They were:

G (group: 1 = ran in place, 2 = did not run in place)

K (smoker: 1 = smokes regularly, 2 = does not smoke regularly)

S (sex: 1 = male, 2 = female)

Letting P_1 = first pulse rate and P_2 = second pulse rate, it was desired to model P_2 as a function of (P_1, G, K, S) . For this a linear model with all possible interactions was considered. The model was of the form

$$\begin{aligned} Y = & \beta_0 + \beta_1 P_1 + \beta_2 G + \beta_3 K + \beta_4 S + \beta_5 GS + \beta_6 GK + \beta_7 SK \\ & + \beta_8 GSK + \beta_9 P_1 G + \beta_{10} P_1 K + \beta_{11} P_1 S + \beta_{12} P_1 GS \\ & + \beta_{13} P_1 GK + \beta_{14} P_1 KS + \beta_{15} P_1 GSK + \varepsilon. \end{aligned} \quad (7.90)$$

Under the assumptions of the GLM the model was fit and the results are shown in Table 7.14, which includes the values of the regression coefficients, their standard errors (S.E.), t -ratios, p -values, variance inflation factors (VIF), and sequential sums of squares (Seq SS).

Table 7.14 Full Model for Pulse Data

Predictor	Coefficient	S.E.	<i>t</i> -statistic	<i>p</i> -value	VIF	Seq SS
constant	149.9	228.5	0.66	0.514	-	-
P_1	-1.577	2.976	-0.53	0.598	1629.6	10096.1
G	-51.6	152.3	-0.34	0.735	8389.8	7908.0
K	-91.0	138.3	-0.66	0.513	6219.3	116.7
S	5.3	168.6	0.03	0.975	10278.6	1087.0
GS	-15.5	124.3	-0.12	0.901	29317.3	2129.0
GK	29.74	88.41	0.34	0.738	16003.6	295.6
SK	30.99	98.14	0.32	0.753	18549.5	62.2
GSK	-4.67	68.55	-0.07	0.946	40712.3	11.1
P_1G	1.032	1.926	0.54	0.594	8712.7	122.4
P_1K	1.402	1.846	0.76	0.450	6878.1	51.9
P_1S	0.504	2.085	0.24	0.810	12649.2	61.8
P_1GS	-0.121	1.494	-0.08	0.936	27402.2	12.6
P_1GK	-0.523	1.150	-0.45	0.651	15125.2	49.6
P_1SK	-0.399	1.242	-0.32	0.749	18958.3	30.1
P_1GSK	0.0772	0.8429	0.09	0.927	35106.2	0.5

Table 7.15 ANOVA for Pulse data

Source	df	Sum of Squares	Mean Squares	<i>F</i>
Regression	15	22034.5	1469.0	24.51
Residual	76	4555.5	59.9	-
Total	91	26590.0	-	-
		$R^2 = 0.829$	$\bar{R}^2 = 0.795$	

Table 7.16 Full Model using 0-1 Dummy Variables

Predictor	Coefficient	S.E.	<i>t</i> -statistic	<i>p</i> -value	VIF	Seq SS
constant	1.25	13.14	0.10	0.925	-	-
P_1	0.9716	0.1839	5.28	0.000	6.2	10096.1
G	16.99	22.91	0.74	0.461	189.9	7908.0
K	9.88	18.13	0.54	0.587	106.8	116.7
S	17.65	19.06	0.93	0.357	131.5	1087.0
GS	24.83	33.40	0.74	0.460	180.2	2129.0
GK	25.06	32.23	0.78	0.439	180.8	295.6
SK	-21.65	55.27	-0.39	0.696	372.3	62.2
GSK	-4.67	68.55	-0.07	0.946	300.0	11.0
P_1G	-0.0196	0.3276	-0.06	0.952	218.2	122.4
P_1K	-0.1115	0.2543	-0.44	0.662	123.5	51.9
P_1S	-0.2264	0.2635	-0.86	0.393	153.7	61.8
P_1GS	-0.0335	0.4512	-0.07	0.941	221.4	12.6
P_1GK	-0.4458	0.4515	-0.99	0.327	211.5	49.6
P_1SK	0.2441	0.6582	0.37	0.712	384.3	30.1
P_1GSK	0.0772	0.8429	0.09	0.927	314.1	0.5

As one can see in Table 7.15 the F -statistic indicates that the variables taken as a whole are significant, but the t values show that none of the variables individually is significant. As we noted in Chapter 5 this behavior is indicative of multicollinearity and that is further substantiated by the high variance inflation factors. At this point it is difficult to interpret what the model is doing.

As a first step towards mitigating the multicollinearity, the qualitative variables were recorded from 1-2 to 0-1. The results are displayed in Table 7.16. The ANOVA is the same as Table 7.15 because a change in the scale does not affect it and now at least P_1 is significant. Moreover, the VIFs are reduced by two orders of magnitude. Again it is difficult to interpret the fit. To further reduce multicollinearity all variables (P_1, G, K, S) were centered

$$P'_1 = P_1 - \overline{P}_1, G' = G - \overline{G}, S' = S - \overline{S}, K = K - \overline{K} \tag{7.91}$$

and the interactions written in terms of the centered variables. Again the model was fit and the results shown in Table 7.17. The effects are dramatic. The variance inflation factors (VIFs) are ≤ 3 (recall $VIF = 1$ for orthogonal variables) and now P_1, G, S and GS are significant. This suggests that the reduced model

$$Y = \beta_0 + \beta_1 P_1 + \beta_2 G + \beta_3 S + \beta_4 GS + \varepsilon \tag{7.92}$$

might be appropriate. A further fit using orthogonalized variables gave results shown in Table 7.18.

Table 7.17 Full Model using Centered Variables

Predictor	Coefficient	S.E.	t -statistic	p -value	VIF	Seq SS
constant	80.919	1.057	76.53	0.000	-	-
P_1	0.81929	0.09264	8.84	0.000	1.6	10096.1
G	-21.935	2.085	-10.52	0.000	1.6	7908.0
K	2.400	2.701	0.89	0.377	2.4	116.7
S	8.604	2.400	3.58	0.001	2.1	1087.0
GS	-22.677	4.664	-4.86	0.000	1.8	2129.0
GK	-7.055	4.998	-1.41	0.162	2.0	295.6
SK	3.492	6.497	0.54	0.592	3.0	62.2
GSK	0.95	11.77	0.08	0.936	2.4	11.0
P_1G	0.1590	0.1887	0.84	0.402	1.6	122.4
P_1K	0.1771	0.2016	0.88	0.382	2.0	51.9
P_1S	-0.1559	0.1927	-0.81	0.421	1.6	61.8
P_1GS	0.0100	0.3814	0.03	0.979	1.7	12.6
P_1GK	-0.4164	0.3893	-1.07	0.288	1.8	49.6
P_1GK	-0.2735	0.4543	-0.60	0.549	2.5	30.1
P_1GSK	0.0772	0.8429	0.09	0.927	2.2	0.5

Table 7.18 Full Model using Orthogonal Variables

Predictor	Coefficient	S.E.	<i>t</i> -statistic	<i>p</i> -value	VIF	Seq SS
constant	43.134	7.516	5.74	0.000	-	-
<i>P</i> ₁	0.76401	0.08419	9.08	0.000	1.3	10096.1
<i>G</i>	-20.846	1.732	-12.04	0.000	1.1	7908.0
<i>K</i>	2.015	1.902	1.06	0.293	1.2	116.7
<i>S</i>	8.358	1.853	4.51	0.000	1.2	1087.0
<i>GS</i>	-22.819	4.108	-5.55	0.000	1.4	2129.0
<i>GK</i>	-7.308	3.797	-1.92	0.058	1.1	295.6
<i>SK</i>	2.208	4.852	0.46	0.650	1.6	62.2
<i>GSK</i>	1.542	9.862	0.16	0.876	1.6	11.0
<i>P</i> ₁ <i>G</i>	0.2262	0.1794	1.26	0.211	1.4	122.4
<i>P</i> ₁ <i>K</i>	0.1848	0.1777	1.04	0.302	1.4	51.9
<i>P</i> ₁ <i>S</i>	-0.1366	0.1826	-0.75	0.457	1.4	61.8
<i>P</i> ₁ <i>GS</i>	0.0113	0.3811	0.03	0.976	1.4	12.6
<i>P</i> ₁ <i>GK</i>	-0.4236	0.3813	-1.11	0.270	1.6	49.6
<i>P</i> ₁ <i>GK</i>	-0.2912	0.4111	-0.71	0.481	1.4	30.1
<i>P</i> ₁ <i>GSK</i>	0.0772	0.8429	0.09	0.927	1.0	0.5

Now the VIFs are all < 2 and five variables are significant. Moreover, now *GK* is marginally significant with a *t* value of -1.92. However, *K* itself does not seem to be significant, indicating the type of anomaly suggested previously. Since logically one would expect the direct effect of a variable to be important if its interaction with another variable is, a general rule is to include all direct effects when their interactions are significant.

As a consequence, a reasonable model would be to include the variables *P*₁, *G*, *K*, *S*, *GS* and *GK* as predictors. To test this, the model was refit using these variables.

Table 7.19 Model using *P*₁, *G*, *K*, *S*, *GS* and *GK*

Predictor	Coefficient	S.E.	<i>t</i> -statistic	<i>p</i> -value	VIF	Seq SS
constant	80.6367	0.8024	100.49	0.000	-	-
<i>P</i> ₁	0.76297	0.07786	9.80	0.000	1.1	10096.1
<i>G</i>	-21.012	1.663	-12.64	0.000	1.0	7908.0
<i>K</i>	8.891	1.769	5.03	0.000	1.2	1087.0
<i>S</i>	1.885	1.786	1.06	0.294	1.1	116.7
<i>GS</i>	-8.030	3.567	-2.25	0.027	1.0	405.6
<i>GK</i>	-20.649	3.510	-5.88	0.000	1.0	2019.0

Now one sees that all variables with the exception of *S* are significant. However since *GS* is significant, *S* should be included for logical completeness. It appears that we can settle for the six variables model as the best choice.

7.6 Logistic Regression Revisited

As indicated in Section 6.6 regression methods can be used to model data with *binary 0-1 responses* as well as *0-1 predictors*. Typically, when one has binary responses, binary predictors occur as well and such models play an important role in interpreting data from medical, social science and engineering experiments. Since models with binary responses

do not have normal errors, one needs to be cautious about using normal theory regression methods without careful examination of the assumptions.

In Section 6.5 we assumed that one might model the frequency of a positive drug response $\pi(\mathbf{x})$ as a linear function of the predictor \mathbf{x} . However, $\pi(\mathbf{x})$ is a probability so it must satisfy $0 \leq \pi(\mathbf{x}) \leq 1$. In general, a linear function $\beta_0 + \sum_{j=1}^m \beta_j x_j$ will not satisfy this condition, so such a model seems to be mathematically inconsistent. To model such data, it appears reasonable to assume that $\langle \beta, \mathbf{x} \rangle = \beta_0 + \sum_{j=1}^m \beta_j x_j$

$$\pi(\mathbf{x}) = \Phi(\langle \beta, \mathbf{x} \rangle) \quad (7.93)$$

for some function Φ such that $0 \leq \Phi(y) \leq 1$, $-\infty < y < \infty$. Examination of experimental data often show that $\pi(\mathbf{x})$ is an S-shaped curve as shown in Figure 7.7.

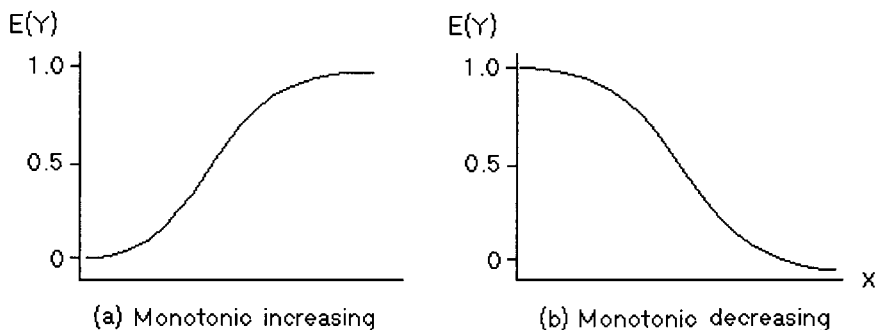


Figure 7.7: Examples of logistic response functions

The S-shape of Φ is typical of a cumulative distribution function (cdf) $F(x)$ of a random variable X . Typical Φ 's are the cdf of a standard normal random variable, referred to as the *probit function* or the *cdf of a logistic random variable*.

$$\Phi(x) = \begin{cases} (1 + e^{-x})^{-1}, & x > 0, \\ 0, & \text{otherwise} \end{cases} \quad (7.94)$$

seems to be the most common choice, particularly in the medical and social sciences. Other choices are cdfs of extreme value and Weibull random variables which are often used in reliability theory. Here we focus on the logistic function, since it is widely used and allows a useful probabilistic interpretation of the coefficients in (7.93).

From (7.94) it follows that by solving

$$\frac{\exp(\langle \beta, \mathbf{x} \rangle)}{1 + \exp(\langle \beta, \mathbf{x} \rangle)} = \pi(\mathbf{x}) \quad (7.95)$$

for $\langle \beta, \mathbf{x} \rangle$ in terms of $\pi(\mathbf{x})$ that

$$\log \left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right] = \langle \beta, \mathbf{x} \rangle \quad (7.96)$$

and this suggests that if Y_i is the relative frequency r_i/n_i of successes when $\mathbf{x} = \mathbf{x}_i$, that a suitable model for binary data is of the form

$$\log \left(\frac{Y_i}{1 - Y_i} \right) = \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle + \varepsilon_i. \quad (7.97)$$

That is we regress

$$\text{logit}(Y_i) \equiv \log \left(\frac{Y_i}{1 - Y_i} \right) \equiv Z_i \quad (7.98)$$

on \mathbf{x}_i . To determine an appropriate estimation method we need to examine the structure of the errors ε_i , $1 \leq i \leq n$, hence of Z_i .

Now if the observations are independent, then r_i , the number of successes in n_i observations, has a binomial distribution with $E(r_i) = n_i \pi(\mathbf{x}_i)$ and $\text{Var}(r_i) = n_i \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]$. Hence,

$$E(Y_i) = E(r_i/n_i) = \pi(\mathbf{x}_i) \quad (7.99)$$

and

$$\text{Var}(Y_i) = \text{Var}(r_i/n_i) = \frac{1}{n_i} \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]. \quad (7.100)$$

When n_i is large, then it follows from the Central Limit Theorem that r_i is approximately $N(n_i \pi(\mathbf{x}_i), n_i \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)])$ and Y_i is approximately $N(\pi(\mathbf{x}_i), \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)] / n_i)$. Using this we can determine the approximate distribution of $\text{logit}(Y_i)$ for large n .

Letting $f(Y_i) = \log[Y_i / (1 - Y_i)]$ it follows from Taylor's theorem that

$$f(Y_i) \simeq f(\mu_i) + (Y_i - \mu_i) f'(\mu_i). \quad (7.101)$$

But, $f(Y) = \log(Y) - \log(1 - Y)$ so that $f'(Y) = 1/Y + 1/(1 - Y)$ and then

$$\begin{aligned} f(Y) &\simeq \log \left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] + [Y_i - \pi(\mathbf{x}_i)] \left[\frac{1}{\pi(\mathbf{x}_i)} + \frac{1}{1 - \pi(\mathbf{x}_i)} \right] \\ &= \text{logit}[\pi(\mathbf{x}_i)] + \frac{Y_i - \pi(\mathbf{x}_i)}{\pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]}. \end{aligned} \quad (7.102)$$

Since Y_i is approximately $N(\pi(\mathbf{x}_i), \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)] / n_i)$, $f(Y_i)$ is approximately a linear transform of a normal random variable so is approximately normal with

$$E[f(Y_i)] \simeq \log \left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] \quad (7.103)$$

and

$$\begin{aligned} \text{Var}[f(Y_i)] &\simeq \frac{1}{\{\pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]\}^2} \text{Var}[Y_i - \pi(\mathbf{x}_i)] \\ &= \frac{1}{\{\pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]\}^2} \frac{\pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]}{n_i} \\ &= \frac{1}{n_i \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]}. \end{aligned} \quad (7.104)$$

Hence, $\text{logit}(Y_i)$ is approximately $N(\text{logit}[\pi(\mathbf{x}_i)], \{n_i \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]\}^{-1})$. From this it follows that the errors ε_i in (7.97) are approximately $N(0, [n_i \pi(\mathbf{x}_i) \{1 - \pi(\mathbf{x}_i)\}]^{-1})$

so our discussion in Section 6.5 suggests that β can be estimated using weighted least squares with

$$\mathbf{W} = \text{diag}(n_i \pi(\mathbf{x}_i) [1 - \pi(\mathbf{x}_i)]). \tag{7.105}$$

Since $\pi(\mathbf{x}_i)$ are unknown, our argument in Section 6.5 suggests a simple approach would be to estimate $\pi(\mathbf{x}_i)$ by $r_i/n_i = f_i$, the observed relative frequency of successes for each covariate combination \mathbf{x}_i . The coefficients β can then be estimated using weighted least squares with weights $w_i = n_i f_i (1 - f_i)$. Further accuracy can generally be achieved using iteratively reweighted least squares now using weights $n_i \hat{\pi}(\mathbf{x}_i) [1 - \hat{\pi}(\mathbf{x}_i)]$ where $\hat{\pi}(\mathbf{x}_i) = \exp(\mathbf{x}_i^T \hat{\beta}) / [1 + \exp(\mathbf{x}_i^T \hat{\beta})]$ where $\hat{\beta}$ is the WLS estimator of β . This can be continued until the estimates of $\hat{\beta}$ converge. Often, the initial WLS estimate is sufficient.

To assess the model, such as goodness of fit, significance of coefficients, etc. One can proceed as indicated in Section 6.5 using the transformed model to obtain t and F tests, residual examination, leverage and influence diagnostics. For further details we refer the reader to [27, 87]. As an example of this approach for analyzing binary response data we consider the following model discussed briefly in [93].

Example 7.8 An experiment was performed to determine customer response to coupons for various levels of price reductions. For this, coupons giving 5%, 10%, 15%, 20% and 30% price reductions were distributed to 200 families ($= n_i$) in each category. To assess the response the number of coupons redeemed was recorded and the results are shown in Table 7.18. Letting r_i = number of coupons redeemed, $f_i = r_i/n_i$ gives the relative frequency of coupon redemption for each category. The results are plotted in Figure 7.8 and the results suggest that a logistic model would be appropriate to model the probability of redemption of $\pi(x_i)$ for a price reduction of $x_i\%$.

Table 7.20 Coupon Redemption Data

Category i	x_i (in %)	n_i	r_i	f_i
1	5	200	32	0.160
2	10	200	51	0.255
3	15	200	70	0.350
4	20	200	103	0.515
5	30	200	148	0.740

The scatter plot of f_i versus x_i suggests that a linear model

$$\text{logit}[\pi(x)] = \beta_0 + \beta_1 x + \varepsilon \tag{7.106}$$

is an appropriate model for the data. To do this the model was fit using a single stage WLS estimation with weights as given in Table 7.21.

Table 7.21 Weights for Coupon Redemption Data

Category i	x_i (in %)	n_i	r_i	f_i	$w_i = n_i f_i (1 - f_i)$
1	5	200	32	0.160	26.88
2	10	200	51	0.255	37.995
3	15	200	70	0.350	45.5
4	20	200	103	0.515	49.955
5	30	200	148	0.740	38.48

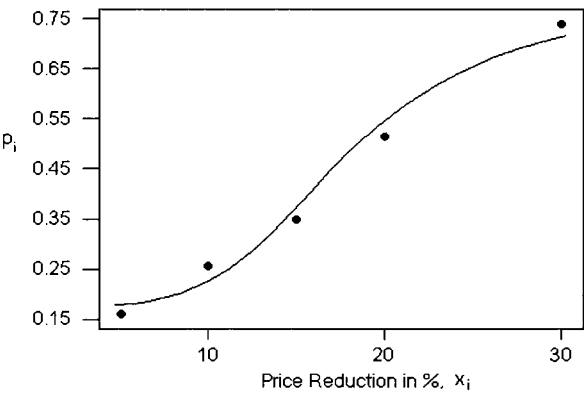


Figure 7.8: Plot of f_i and fitted Logistic response function

The results of the fit are shown in Table 7.22 with the fitted equation being

$$\text{logit} [\pi (x)] = -2.19 + 0.109x. \tag{7.107}$$

Table 7.22 Analysis of Parameter Estimates

Predictor	Coefficient	S.E. Coeff.	<i>t</i> -statistic	<i>p</i> -value
constant	−2.18506	0.06783	−32.21	0.000
Price	0.10870	0.00363	29.93	0.000

Table 7.23 ANOVA Table for Coupon Redemption Data

Source	df	Sum of Squares	Mean Squares	<i>F</i>
Regression	1	151.98	151.98	895.53
Residual	3	0.51	0.17	
Total	4	152.49		

The t values show that both β_0 and β_1 are highly significant. Using (7.107) the estimated values $\hat{\pi}(x)$ are shown in Table 7.24 and a plot of the residuals is given in Figure 7.9. Overall, it appears that the model (7.107) gives an excellent representation of the observed data.

Table 7.24 $\hat{\pi}(x)$ for Coupon Redemption Data

Category i	x_i (in %)	n_i	r_i	f_i	$\text{logit}[\hat{\pi}(x)]$
1	5	200	32	0.160	−1.645
2	10	200	51	0.255	−1.10
3	15	200	70	0.350	−0.555
4	20	200	103	0.515	−0.01
5	30	200	148	0.740	1.08

7.6.1 Interpretation of Logistic Coefficients

One of the attractive features of the logistic model is the probabilistic interpretation of the regression coefficients that is possible, particularly when the independent variable

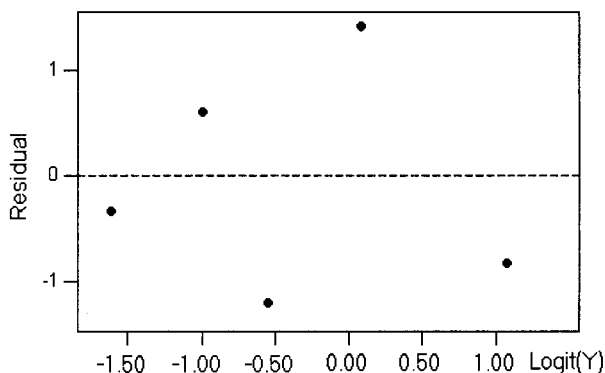


Figure 7.9: Scatter plot of residuals versus logit(Y)

is categorical with two values 0-1. As we have already discussed, these are useful in describing treatment effects or the presence or absence of a factor, such as sex, smoking etc.

From (7.106) a change of one unit in the independent variable is given by

$$\begin{aligned}\beta_i &= \text{logit}[\pi(x_i = 1)] - \text{logit}[\pi(x_i = 0)] \\ &= \log \frac{\{\pi(x_i = 1) / [1 - \pi(x_i = 1)]\}}{\{\pi(x_i = 0) / [1 - \pi(x_i = 0)]\}}.\end{aligned}\quad (7.108)$$

Now if π represents the probability of an event, then $\pi / (1 - \pi)$ is the *odds* of the event occurring against its not occurring. Hence, $\text{logit}(\pi)$ is often referred to as the *log odds* of the event. Using this terminology, the quantity

$$\frac{\pi(x_i = 1) / [1 - \pi(x_i = 1)]}{\pi(x_i = 0) / [1 - \pi(x_i = 0)]} \quad (7.109)$$

is called the *odds ratio* and the (7.108) is then the *log odds ratio*. If we now exponentiate (7.108) we get $\exp(\beta_i) \equiv \psi_i$. For example, if $x_i = 1$ for a treatment and $x_i = 0$ otherwise, then ψ_i represents the odds ratio for the success of the treatment against its failure. In many situations, where the absolute risk π of an event is small, such as getting cancer from an air pollutant, then $1 - \pi \simeq 1$ and the odds ratio

$$\psi_i \simeq \frac{\pi(x_i = 1)}{\pi(x_i = 0)} \quad (7.110)$$

where $\pi(x_i = 1) / \pi(x_i = 0)$ is called the *relative risk* of the event of getting cancer due to the presence of a pollutant ($x_i = 1$) against its absence ($x_i = 0$). So if $\psi = 2$, then we would interpret this as saying one is twice as likely as getting cancer from exposure to the pollutant against not being exposed. It is the possibility of making such interpretations which contributes to the use of the logistic model. Such quantities are reported almost daily in the popular media.

7.6.2 Maximum Likelihood Estimation

When the sample sizes n_i in (7.98) are not large (in practice they may be zero) then the asymptotic approach to estimation of the parameters in a logistic model may not be appropriate. In this case an estimation method which does not require large sample sizes should be considered. In keeping with our treatment of the GLM this can be done using maximum likelihood estimation using grouped or ungrouped data. We will consider the case of ungrouped data, since weighted least squares cannot be used directly in this case ($\text{logit}(Y)$, $Y = 0, 1$ is undefined).

Hence, we assume that we have n independent observations of a random variable Y with binary outcomes 0,1. Letting Y_i be the outcome of the i -th observation, then Y_i is a Bernoulli random variable with $P\{Y_i = 1\} = \pi(\mathbf{x}_i) \equiv \pi_i$ where \mathbf{x}_i is the covariate vector for the i -th observation. Then it follows from (2.143) that the likelihood function for Y_i is

$$f_{Y_i}(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (7.111)$$

so the likelihood function for all n observations is

$$L = \prod_{i=1}^n f_{Y_i}(y_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}. \quad (7.112)$$

If we assume that π_i is given by the logistic function

$$\pi_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \quad (7.113)$$

then $\boldsymbol{\beta}$ can be found by setting

$$\frac{\partial L}{\partial \beta_j} = 0, \quad 0 \leq j \leq m. \quad (7.114)$$

As for normal random variables this is done more conveniently by setting

$$\frac{\partial}{\partial \beta_j} \log L = 0, \quad 0 \leq j \leq m. \quad (7.115)$$

Now

$$\mathcal{L} = \log L = \sum_{i=1}^n y_i \log \pi_i + (1 - y_i) \log (1 - \pi_i) \quad (7.116)$$

and

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^n y_i \frac{\partial}{\partial \beta_j} \log \pi_i + (1 - y_i) \frac{\partial}{\partial \beta_j} \log (1 - \pi_i). \quad (7.117)$$

Since

$$\frac{\partial \pi_i}{\partial \beta_0} = -\pi_i(1 - \pi_i) \quad \text{and} \quad \frac{\partial \pi_i}{\partial \beta_j} = -x_{ij}\pi_i(1 - \pi_i), \quad 0 \leq j \leq m \quad (7.118)$$

it follows using the chain-rule from calculus that

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^n y_i \left[\frac{-\pi_i(1 - \pi_i)}{\pi_i} \right] + \sum_{i=1}^n (1 - y_i) \left[\frac{\pi_i(1 - \pi_i)}{1 - \pi_i} \right] \quad (7.119)$$

and

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^n y_i x_{ij} \left[\frac{-\pi_i (1 - \pi_i)}{\pi_i} \right] + \sum_{i=1}^n (1 - y_i) x_{ij} \left[\frac{\pi_i (1 - \pi_i)}{1 - \pi_i} \right]. \quad (7.120)$$

Thus, (7.118)-(7.120) give

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^n -y_i (1 - \pi_i) + (1 - y_i) \pi_i = \sum_{i=1}^n (-y_i + \pi_i) = 0 \quad (7.121)$$

and

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_j} &= \sum_{i=1}^n y_i x_{ij} (1 - \pi_i) + \sum_{i=1}^n (1 - y_i) x_{ij} \pi_i \\ &= \sum_{i=1}^n (-y_i x_{ij} + x_{ij} \pi_i) = 0, \quad 1 \leq j \leq m. \end{aligned} \quad (7.122)$$

Hence, the MLE equations for β are given by

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \pi_i \quad (7.123)$$

and

$$\sum_{i=1}^n x_{ij} (\pi_i - y_i) = 0, \quad 1 \leq j \leq m. \quad (7.124)$$

Since π_i depends nonlinearly on the parameters, Equations (7.121)-(7.122) represent $m + 1$ simultaneous nonlinear equations which generally must be solved numerically in some fashion. Usually, this is done by a Newton iteration scheme which can be shown to be equivalent to an iteratively reweighted least squares method analogous to that discussed previously for grouped data. For large sample sizes the asymptotic theory is essentially equivalent to the WLS estimation given in Section 6.7 [64]. So inference, goodness of fit and diagnostics can again be based on those of the theory of the generalized regression model. We refer the reader to [27, 87] for further details.

7.7 The Generalized Linear Model

Although the general linear model with normal errors has been shown to provide an adequate method for modeling a wide variety of statistical data, we have already seen that there are many situations where the normal error model is inappropriate. For example, we showed in Chapter 6 that the drink delivery data was fit better by a power family model of the Box-Cox type and binary response data was generally better explained using logistic regression. Other types of data clearly have non-normal error distributions. These include random count data which often follow a Poisson distribution and survival data which often have an exponential or gamma distribution. In this section we briefly discuss a class of models which include all of these as particular cases, the *generalized linear model* (GLIM) first introduced by Nelder and Wedderburn in 1972 [92].

The basic idea is to obtain a *linear predictor*, $\mathbf{x}_i^T \boldsymbol{\beta}$ as a function of the mean response as a way of combining the simplicity of the linear model with the generality of a large-family of non-normal error distributions described by the exponential family of random variables. The basic idea is the concept of a link function which connects the mean response to a given error distribution.

7.7.1 Linear Predictors and Link Functions

Let $Y_i, i = 1, 2, \dots, n$ represent the outcome of the i -th random observation. We assume that $Y_i, i = 1, 2, \dots, n$ belong to the same family of random variables which differ only in their means

$$\mu_i = E(Y_i), \quad i = 1, 2, \dots, n. \quad (7.125)$$

As for the GLM and logistic models, we assume that there exists a function g such that

$$g(\mu_i) = \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (7.126)$$

where \mathbf{x}_i is an $m + 1$ vector of response variable and $\boldsymbol{\beta}$ is an $m + 1$ vector of unknown coefficients which have to be estimated from the data. The function g is referred to as the *link function*. For example, in the GLM $g(\mu) = \mu$, the *identity link*, while for logistic regression

$$g(\mu_i) = g(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \equiv \text{logit}(\pi_i). \quad (7.127)$$

If g is invertible, then (7.126) gives

$$\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}). \quad (7.128)$$

For the GLM, g^{-1} is the identity, while for logistic regression

$$g^{-1}(\pi) = \frac{\exp(\pi)}{1 + \exp(\pi)} \quad (7.129)$$

so

$$g^{-1}(\mu_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}. \quad (7.130)$$

As we have observed in dealing with binary and binomial response data, link functions may often be taken as inverse cdfs. For example,

(a) the *probit link*,

$$g(\mu) = \Phi^{-1}(\mu); \quad (7.131)$$

where Φ is the cdf of a $N(0, 1)$ random variable;

(b) the *complementary log-log link*,

$$g(\mu) = \log[\log(1 - \mu)]; \quad (7.132)$$

and

(c) the *power family*,

$$g(\mu) = \begin{cases} \mu^\lambda, & \lambda \neq 0, \\ \log \mu, & \lambda = 0. \end{cases} \quad (7.133)$$

Other link functions, so-called *canonical links* are given in Table 7.23.

Table 7.23 Canonical Links and Error Functions in GLIM

Distribution of Error Function	Canonical Link Function η_i	Name of Link	Regression Model (Inverse Link Function)
(1) Normal	$\eta_i = \mu_i$	Identity	$E(Y) = \mathbf{x}^T \boldsymbol{\beta}$
(2) Binomial	$\eta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right)$	Logistic	$E(Y) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T \boldsymbol{\beta})}$
(3) Poisson	$\eta_i = \log(\lambda_i)$	Logarithm	$E(Y) = \exp(\mathbf{x}^T \boldsymbol{\beta})$
(4) Exponential	$\eta_i = 1/\lambda_i$	Reciprocal	$E(Y) = (\mathbf{x}^T \boldsymbol{\beta})^{-1}$
(5) Gamma	$\eta_i = 1/\lambda_i$	Reciprocal	$E(Y) = (\mathbf{x}^T \boldsymbol{\beta})^{-1}$

7.7.2 The Error Function

The question now arises as to what is the relation between the link function and the underlying distribution of Y_i . In practice, what is the role of the canonical links given in Table 7.23. This is best explained through the use of the exponential family.

Recall from Eq. (2.37) that Y is said to belong to the exponential family if

$$f_Y(y, \theta) = \exp[a(x)b(\theta) + c(\theta) + d(x)]. \quad (7.134)$$

As we have already seen, normal, binomial and Poisson random variables are all members of the exponential family and when $a(x) \equiv x$ and $b(\theta) = \psi$, then (7.134) can be written in the *canonical form*

$$f_X(x, \psi) = \exp[x\psi + b'(\psi) + d(x)] \quad (7.135)$$

If $b(\theta)$ is a function of $E(Y) = \mu$, then letting

$$b(\mu) = \psi = \langle \boldsymbol{\beta}, \mathbf{x} \rangle = \mathbf{x}^T \boldsymbol{\beta}$$

b is the canonical link. We examine this for a member of cases given in Table 7.23.

First, if $Y \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} f(y, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (x^2 - 2\mu x + \mu^2) \right]. \end{aligned} \quad (7.136)$$

Some simple algebra shows that the canonical link

$$b(\mu) = \mu. \quad (7.137)$$

If Y is a Bernoulli random variable, then

$$\begin{aligned} f(y, p) &= p^x (1-p)^{1-x} \\ &= \exp(x \log p) \exp[(1-x) \log(1-p)] \\ &= \exp[x \log p - x \log(1-p) + \log(1-p)] \\ &= \exp\{x \log[p/(1-p)] + \log(1-p)\}. \end{aligned} \quad (7.138)$$

Since $E(Y) = p = \mu$, then (7.138) can be written in the form

$$f(x, \mu) = \exp [x \log [\mu / (1 - \mu)] + \log (1 - \mu)]. \quad (7.139)$$

Letting

$$b(\mu) = \log [p / (1 - p)] = \psi = \langle \beta, \mathbf{x} \rangle \quad (7.140)$$

we arrive at the link function for the logistic model described in the previous two sections.

Last, we consider Y having an exponential distribution. Then the density of this given by

$$\begin{aligned} f(x, \mu) &= \mu^{-1} \exp(-x/\mu) \\ &= \exp(-x/\mu - \log \mu), \end{aligned} \quad (7.141)$$

where $E(Y) = \mu$. Hence, the canonical link

$$b(\mu) = -1/\mu = \psi = \langle \beta, \mathbf{x} \rangle. \quad (7.142)$$

So the canonical link is $-1/\mu$. Since the sign is irrelevant, we can take the canonical link as

$$b(\mu) = 1/\mu \quad (7.143)$$

as given in Table 7.23. The remaining entries in the table can be obtained in the same way.

7.7.3 Parameter Estimation

As for the GLM and logistic model, the basic problem in GLIM modeling is the estimation of the parameters in the relation.

$$\mu_i = g^{-1}(\langle \mathbf{x}_i, \beta \rangle) = g^{-1}(\mathbf{x}_i^T \beta) \quad (7.144)$$

where $\mathbf{x}_i, i = 1, 2, \dots, n$ are n observations of the m predictor variables. As for the GLM and logistic regression this is most commonly done by maximum likelihood estimation.

Writing the density of Y_i in canonical form

$$f(y_i, \beta) = \exp [y_i b^{-1}(\mathbf{x}_i^T \beta) + c'(\mathbf{x}_i^T \beta) + d(y_i)] \quad (7.145)$$

so that

$$\log [f(y_i, \beta)] = y_i b^{-1}(\mathbf{x}_i^T \beta) + c'(\mathbf{x}_i^T \beta) + d(y_i). \quad (7.146)$$

Then the log-likelihood function \mathcal{L} of the n observations $y_i, i = 1, 2, \dots, n$ is given by

$$\mathcal{L} = \sum_{i=1}^n [y_i b^{-1}(\mathbf{x}_i^T \beta) + c'(\mathbf{x}_i^T \beta) + d(y_i)] \quad (7.147)$$

The MLE of β is then obtained by setting

$$\frac{\partial \mathcal{L}}{\partial \beta} = 0. \quad (7.148)$$

This generally leads to a set of nonlinear equations for β which usually must be solved by iteration [27, 87]. If Newton's method is used as the numerical iterative method it can

be shown that this can be done by solving a sequence of *weighted least squares equations* of the form

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{W} \mathbf{z}, \tag{7.149}$$

where $\mathbf{X} = [x_{ij}]$, $1 \leq i \leq n$, $1 \leq j \leq m$ is the design matrix and \mathbf{W} and \mathbf{z} generally depend on $\boldsymbol{\beta}$. Details can be found in [85, 25, 52].

Inferences concerning the model's correctness can be made via generalized likelihood ratio tests. For example, if $\mathcal{L}(\hat{\boldsymbol{\beta}})$ is the maximized log-likelihood with respect to $\boldsymbol{\beta}$ and $\mathcal{L}(\hat{\boldsymbol{\mu}})$ is the log-likelihood for the null model we then can obtain the goodness of fit measure the *deviance*

$$\mathcal{D}(\hat{\boldsymbol{\beta}}) = -2 \left[\mathcal{L}(\hat{\boldsymbol{\beta}}) - \mathcal{L}(\hat{\boldsymbol{\mu}}) \right] \tag{7.150}$$

which asymptotically has a χ^2 -distribution with $n - m - 1$ degrees of freedom under the null hypothesis H_0 : the model being fit is correct. Hence, large values of \mathcal{D} are taken to be indicative of a poor fit for the $\langle \mathbf{x}, \boldsymbol{\beta} \rangle$ model. If $\mathcal{D}(\hat{\boldsymbol{\beta}}) > \chi^2(n - m - 1, \alpha)$ we conclude there is a lack of fit at the α level of significance and reject H_0 . We note that differences in deviances may be used in a way analogous to the extra sums of squares in the GLM.

For more details we refer readers to [85, 25, 52].

7.8 Exercises

7.1 A set of data [*Journal of Pharmaceutical Sciences* (1991) **80**, pp. 971-977] was obtained on the observed mole fraction solubility of a solute at a constant temperature. The response Y is the negative logarithm of the mole fraction solubility, along with

- x_1 = dispersion partial solubility,
- x_2 = dipolar partial solubility,
- x_3 = hydrogen bonding Hansen partial solubility.

Answer the following questions using the data in Table 7.24.

Table 7.24 Mole Fraction Solubility Data

No.	y	x_1	x_2	x_3	No.	y	x_1	x_2	x_3
1	0.2220	7.3	0.0	0.0	14	0.1010	7.3	2.5	6.8
2	0.3950	8.7	0.0	0.3	15	0.2320	8.5	2.0	6.6
3	0.4220	8.8	0.7	1.0	16	0.3060	9.5	2.5	5.0
4	0.4370	8.1	4.0	0.2	17	0.0923	7.4	2.8	7.8
5	0.4280	9.0	0.5	1.0	18	0.1160	7.8	2.8	7.7
6	0.4670	8.7	1.5	2.8	19	0.0764	7.7	3.0	8.0
7	0.4440	9.3	2.1	1.0	20	0.4390	10.3	1.7	4.2
8	0.3780	7.6	5.1	3.4	21	0.0944	7.8	3.3	8.5
9	0.4940	10.0	0.0	0.3	22	0.1170	7.1	3.9	6.6
10	0.4560	8.4	3.7	4.1	23	0.0726	7.7	4.3	9.5
11	0.4520	9.3	3.6	2.0	24	0.0412	7.4	6.0	10.9
12	0.1120	7.7	2.8	7.1	25	0.2510	7.3	2.0	5.2
13	0.4320	9.8	4.2	2.0	26	0.0 ⁴²	7.6	7.8	20.7

- (a) Fit a complete second-degree polynomial model to the data.
- (b) Test for significance of the regression, and construct t statistics for each model parameter. Take $\alpha = 0.1$. Interpret these results.
- (c) Plot the residuals and comment on model adequacy.
- (d) Use the extra-sum-of-squares method to test the contribution of all second-order terms to the model. Take $\alpha = 0.05$.

7.2 A scientist collected experimental data [90] on the radius of a propellant grain (y) as a function of powder temperature, x_1 , extrusion rate, x_2 , and die temperature, x_3 .

Grain Radius (y)	Powder Temperature (x_1)	Extrusion Rate (x_2)	Die Temperature (x_3)
82	150	12	220
93	190	12	220
114	150	24	220
124	150	12	250
111	190	24	220
129	190	12	250
157	150	24	250
164	190	24	250

- (a) Consider the multiple linear regression model with centered regressors

$$y_i = \beta_0^* + \beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + \beta_3 (x_{3i} - \bar{x}_3) + \varepsilon_i.$$

Write the observation vector \mathbf{y} , the design matrix \mathbf{X} , and the parameter vector $\boldsymbol{\beta}$ in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- (b) Write out the normal equations for least-squares estimation

$$(\mathbf{X}^T \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}.$$

- (c) What characteristic in this experiment do you suppose produces the special form of $\mathbf{X}^T \mathbf{X}$ in (b)?

- (d) Estimate the regression coefficients in the model in (a).

- (e) Test the hypothesis $H_0 : \beta_1 = 0$ and $\beta_2 = 0$.

- (f) Find the hat matrix \mathbf{H} .

- (h) Find the VIFs (variance inflation factors) of the coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. Do you have any explanation as to why these measures of damage due to collinearity give the results that they do?

7.3 Bars of soap are scored for their appearance in a manufacturing operation. These scores are on a 1-10 scale, and the higher the score the better. The difference between operator performance and the speed of the manufacturing line are believed to

measurably affect the quality of the appearance. The following data were collected on this problem:

Operator No.	Line Speed	Appearance (Sum for 30 Bars)
1	150	255
1	175	246
1	200	249
2	150	260
2	175	223
2	200	231
3	150	265
3	175	247
3	200	256

- (a) Using dummy variables, specify the design and parameter matrices.
 - (b) Fit a multiple regression model to these data.
 - (c) Construct the ANOVA table. Using $\alpha = 0.05$, determine whether operator differences are important in bar appearance.
 - (d) Does line speed affect appearance? Take $\alpha = 0.05$.
 - (e) Using the regression model, show that the average appearance score for operator #1 is 250, operator #2 is 238, and operator #3 is 256.
 - (f) Plot the residuals. What model would you use to predict bar appearance?
- 7.4 A set of 5 pairs of observations were obtained as below. Using the method of orthogonal polynomials described in Section 7.2 answer the questions below.

y (index)	9.8	11.0	13.2	15.1	16.0
x (year)	1980	1981	1982	1983	1984

- (a) Fit a third-degree equation to the above data.
 - (b) Test the hypothesis that a second-degree equation is adequate. Take $\alpha = 0.05$.
- 7.5 An experimenter suggests the following dummy variable schemes to separate how possible levels depend upon groups. Are they permissible?

(a)	Z_1	Z_2	Z_3					
	1	-1	-1					
	0	2	-2					
	0	0	3					
(c)	Z_1	Z_2	Z_3	Z_4				
	1	-1	-1	-1				
	1	-1	-1	1				
	1	-1	1	-1				
	-1	-1	1	1				
	-1	1	-1	-1				
(b)	Z_0	Z_1	Z_2	Z_3				
	1	-1	-1	-1				
	1	1	0	0				
	1	0	1	0				
	1	0	0	1				
(d)	Z_0	Z_1	Z_2	Z_3	Z_4			
	1	1	-1	-1	-1			
	1	-1	2	-1	-1			
	1	-1	-1	3	-1			
	1	-1	-1	-1	4			
	1	-1	-1	-1	-1			

7.6 Consider the data for Exercise 6.3.

(a) Fit a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i.$$

(b) Compute the ordinary residuals, the PRESS residuals, and the sum of the absolute PRESS residuals.

7.7 An experiment was conducted in the civil engineering department at Virginia Polytechnic Institute and State University in 1988 to investigate the growth of a certain type of algae in water. A set of observations was obtained as a function of time, denoted by x_2 (days), the dosage (in mg) of copper added to the water, denoted by x_1 , and y denotes the units of algae observed. The data are shown in Table 7.25.

Table 7.25 Algae Growth and Copper Data

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
.30	1	5	.37	1	12	.23	1	18	.36	1	25
.34	1	5	.36	1	12	.23	1	18	.36	1	25
.20	2	5	.30	2	12	.28	2	18	.24	2	25
.24	2	5	.31	2	12	.27	2	18	.27	2	25
.24	2	5	.30	2	12	.25	2	18	.31	2	25
.28	3	5	.30	3	12	.27	3	18	.26	3	25
.20	3	5	.30	3	12	.25	3	18	.26	3	25
.24	3	5	.30	3	12	.25	3	18	.28	3	25
.02	4	5	.14	4	12	.06	4	18	.14	4	25
.02	4	5	.14	4	12	.10	4	18	.11	4	25
.06	4	5	.14	4	12	.10	4	18	.11	4	25
0	5	5	.14	5	12	.02	5	18	.04	5	25
0	5	5	.15	5	12	.02	5	18	.07	5	25
0	5	5	.15	5	12	.02	5	18	.05	5	25

(a) Fit the data to the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i.$$

(b) Test the interaction term: $H_0 : \beta_{12} = 0$ versus $H_0 : \beta_{12} \neq 0$.

(c) Make a test for the lack-of-fit and draw a conclusion.

(d) Draw the scatter plots of the residuals for the fitted model against x_1 and x_2 separately. Give a comment on each of them.

7.8 Suppose that you have two sets of data with pairs of values (x, y) . You consider the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + Z (\gamma_0 + \gamma_1 x + \gamma_{11} x^2) + \varepsilon,$$

using a dummy variable Z whose value is -1 for set A and 1 for set B because you are not sure whether to fit the data separately or together.

(a) Set up a hypothesis for the case of fitting a single quadratic model for all the data.

(b) Set up a hypothesis for the case of fitting a single linear model for all the data.

(c) How would you obtain separate quadratic fits to the two data sets?

7.9 Consider the data below [90]. Varying numbers of fabric type specimens were exposed to load (x) in lb/in². Also listed is the number of specimens that failed.

Load (x)	No. Specimens	No. of Failures
5	600	13
35	500	95
70	600	189
80	300	95
90	300	130

- (a) Fit the data using a logistic model.
- (b) Find the maximum likelihood estimates of β_0, β_1 .
- (c) Using the fitted values compare the two models: the linear model and the logistic model.
- 7.10** In an experiment testing the effect of a toxic substance, 1,500 experimental insects were divided randomly into six groups of 250 each. After the insects in each group were exposed to a fixed dose of the toxic substance, the number of dead insects was counted. The results are shown below where $\log(x_j)$ denotes the dose level on a logarithmic scale in group j and R_j represents the number of dead insects in the j -th group. Assume that the logistic response function is appropriate [93].

Group j	1	2	3	4	5	6
$\log(x_j)$	1	2	3	4	5	6
n_j	250	250	250	250	250	250
R_j	28	53	93	126	172	197

- (a) Find the maximum likelihood estimates of β_0, β_1 . Write down the fitted logistic function.
- (b) Plot the fitted response function and the estimated proportions p_j on the same graph. Give a comment.
- (c) What is the estimated probability that an insect dies when the dose level is $x = 3.5$?
- (d) Find the estimated median lethal dose - that is, the dose for which 50 percent of the experimental insects are expected to die.
- 7.11** Indicate how you would use the method of least squares to fit a model of the form

$$y = \gamma_0 x^{\gamma_1}$$

where both γ_0 and γ_1 are unknown.

7.12 The population (in millions) of the United States from 1790 to 1970 is given in Table 7.26:

Table 7.26 Population of the U.S. 1790-1970

Year	Population	Year	Population
1790	3.929	1890	62.9
1800	5.308	1900	76.0
1810	7.240	1910	92.0
1820	9.640	1920	105.7
1830	12.870	1930	122.8
1840	17.070	1940	131.7
1850	23.190	1950	151.3
1860	31.440	1960	179.3
1870	39.820	1970	203.2
1880	50.160		

- (a) Fit a model of the form $y = \beta_0 \exp(\beta_1 x)$ to these data using the method of least squares.
- (b) Use your model found in (a) to predict the population in 1980 and compare it with the true value.
- (c) Use only the data from 1790 to 1900 to fit the model in (a).
- (d) If you were a demographer in 1900 and were asked to predict the population for the next century on the basis of the model in (c), do you think you would have done a good job? (Since an exponential model implies a constant growth rate, such models tend to be inappropriate over long periods of time. A more realistic model is the logistic regression model.)
- (e) Apply the logistic regression model to these data.

7.13 Consider a regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where ϵ is $N(0, \sigma_0^2 \mathbf{I})$, σ_0 is known and $\mathbf{X} = [\mathbf{X}_1 : \mathbf{X}_2]$ where \mathbf{X}_1 is $n \times r$ and \mathbf{X}_2 is $n \times (p - r)$. Suppose one wishes to test

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

- (a) Consider a likelihood ratio test for these hypotheses. Write out the statistic

$$\lambda(\beta_1 | \beta_2) = -2 \log \left[L(\hat{\beta}_2) / L(\hat{\beta}) \right]$$

where $L(\hat{\beta}_2)$ is the maximized likelihood under the restricted model and $L(\hat{\beta})$ is the maximized likelihood under the full model.

- (b) How is this statistic related to the difference in the SSE (residual sum of squares) for the full and reduced models?

7.14 Suppose that the log-link is used, producing the model

$$\mu_i = E(Y_i) = \exp(\mathbf{x}_i^T \beta), i = 1, 2, \dots, n.$$

Show that the maximum likelihood procedure for estimation of β results in the solution to

$$\mathbf{X}^T \hat{\boldsymbol{\varepsilon}} = \mathbf{0}$$

where $\hat{\boldsymbol{\varepsilon}}$ is the residual vector

$$\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} y_1 - \hat{\mu}_1 \\ y_2 - \hat{\mu}_2 \\ \vdots \\ y_n - \hat{\mu}_n \end{pmatrix}.$$

[Hint: The estimation is given as the solution to $\hat{\beta}$ in

$$\mathbf{X}^T (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0} \text{ where } \hat{y}_i = \exp(\mathbf{x}_i^T \beta).$$

In other words, you are required to find the solution to $\hat{\beta}$ of $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$. So you need merely to maximize the log likelihood for this case and show that this leads you to the above.]

7.15 Refer to Table 7.23, for canonical links and error functions in GLIM.

(a) Write the gamma density in the form of exponential family,

$$f(y_i; \theta_i) = \exp[a(y_i)b(\theta_i) + g(\theta_i) + h(y_i)].$$

(b) Show how the canonical link function (and thus the regression model) is developed for the case of the gamma distribution.

[Hint: Using the result in (a), $g(\theta_i)$ and μ_i . Then write $\theta_i = \mathbf{x}_i^T \beta$, and thus $\mathbf{x}_i^T \beta = g(\mu_i)$.]

7.16 Suppose that a nonlinear regression model $y = f(\theta, x) + \varepsilon$ is considered with

$$f(x) = \exp(\theta x).$$

The following observations are given:

Obs. No.	x	y
1	1	2.713
2	2	3.025
3	3	11.731

(a) Write down the normal equations of the least squares method to estimate the parameter θ .

(b) Using an iterative method, determine the value of the estimator $\hat{\theta}_{LSE}$ in (a).

(c) Find the MLE of θ with $\varepsilon \sim N(0, \sigma^2)$. Does $\hat{\theta}_{MLE}$ coincide with $\hat{\theta}_{LSE}$ in (a)?

Chapter 8

Selection of a Regression Model

8.1 Introduction

In the preceding chapters, we have discussed how to test for the significance of the postulated model through the (i) the lack of fit test (ii) examination of residuals and (iii) checking of normality assumptions. These exploratory techniques often lead us to alterations in the initial model such as transformations of the data or further regression techniques. However, in many practical situations, we will be looking for the most appropriate subset of regressors or predictor variables that should be used in the model. Some aspects of this issue have already been examined in Chapters 5 and 6 using the above mentioned exploratory methods. This is called the *variable selection problem* or equivalently *selecting the best subset*. Hence, the main purpose of this chapter is to provide methods and criteria for doing this [57, 90, 113, 114].

Using a model different from the true one is called *model misspecification*. As we shall see, there are two major consequences of model misspecification:

1. If some variables are omitted in the model, then the estimates of the remaining variables are biased.
2. If there are too many variables, then in general the variances of the remaining variables become large.

In some sense, model selection is a trade-off between biasedness and precision. Therefore, the main idea for selecting an appropriate/best regression model is to find a compromise between two conflicting criteria: (a) *reliability* in prediction and (b) *parsimony* (simplicity) of model specification from both the economical and practical points of view. Since the term “best” is somewhat subjective, the ideal model should include the fewest number of regressors that permit an adequate prediction (or interpretation) of the responses. Usually, a unique best regression model does not exist, nor is there a unique statistical procedure for choosing the subset; certain professional or personal judgement is needed for all the methods described in this chapter. For instance, if two regressor variables are highly correlated with Y (response variable) and highly correlated with each

other, then it is often sufficient to include just one of the regressors in the model. The choice of which regressor variable to include may depend, for example, on which variable is easier to measure or cheaper to obtain.

Thus, in this chapter we shall examine in detail both criteria functions and computational techniques for finding the best subset in a regression model.

8.2 Consequences of Model Misspecification

Before examining a number of criteria for model selection, we discuss a number of consequences of omitting variables from the true, but generally unknown full model. These will further motivate various methods for model selection. We first consider the effects on the estimates of the regression coefficients β .

Again we suppose the full model can be written as in (5.16)

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon. \quad (8.1)$$

Assume now that β is partitioned as

$$\beta = (\beta_1, \beta_2)^T \quad (8.2)$$

where $\beta_1 = (\beta_0, \beta_1, \dots, \beta_p)^T$, $\beta_2 = (\beta_{p+1}, \beta_{p+2}, \dots, \beta_m)^T$. β_1 is the vector of coefficients in the reduced model and β_2 is the vector of variables that are deleted (note that this may require a permutation of the regressors). Also note that $\dim(\beta_1) = p + 1$ and $\dim(\beta_2) = m + 1 - (p + 1) = m - p$. Then \mathbf{X} can be partitioned conformally as

$$\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2] \quad (8.3)$$

so that (8.1) can be written as

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon \quad (8.4)$$

and the reduced model is given by

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \varepsilon. \quad (8.5)$$

From (5.22b) the least squares estimate of β is given by

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (8.6)$$

and similarly the least squares estimate of β_1 is given by

$$\hat{\beta}_1 = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{Y}. \quad (8.7)$$

Recall from Chapter 5 that $\hat{\beta}$ is an unbiased estimate of β . However, $\hat{\beta}_1$ is not an unbiased estimate of β_1 . To see this note that

$$E(\hat{\beta}_1) = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T E(\mathbf{Y}). \quad (8.8)$$

From (8.4) $E(\mathbf{Y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ and using this in (8.7) gives

$$\begin{aligned} E(\hat{\beta}_1) &= (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T (\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2) \\ &= (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_1\beta_1 + (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2\beta_2 \end{aligned} \quad (8.9)$$

where $\mathbf{A} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2$ is called the *alias* or *bias matrix*. In general, $\mathbf{A} \neq \mathbf{0}$ so that $E(\hat{\beta}_1) \neq \beta_1$. Hence, generally $\hat{\beta}_1$ is biased. However, in the special case where the columns of \mathbf{X} are orthogonal, then $\mathbf{X}_1^T \mathbf{X}_2 = 0$, and $\hat{\beta}_1$ is unbiased.

The effect on the variance of $\hat{\beta}_1$ is somewhat more complicated and we refer the reader to the paper of Hocking [57] for full details. Recall again from Theorem 5.4 that

$$\Sigma(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (8.10)$$

and

$$\Sigma(\hat{\beta}_1) = \sigma^2 (\mathbf{X}_1^T \mathbf{X}_1)^{-1}. \quad (8.11)$$

It can be shown that if $\hat{\beta}_1^*$ is the least squares estimator of β_1 in $\beta = (\beta_1 | \beta_2)^T$ from the full model, that

$$\Sigma(\hat{\beta}_1^*) - \Sigma(\hat{\beta}_1) \quad (8.12)$$

is positive semi-definite so that the variances of $\hat{\beta}_1$ are at least as large as those of $\hat{\beta}_1^*$. In general, the variances of $\hat{\beta}_1^*$ will be larger than those of $\hat{\beta}_1$. Hence, deleting variables increases the precision of the estimates of $(\beta_0, \beta_1, \dots, \beta_p)^T$.

It is also useful to examine the effect of deleting variables on the predicted values. Letting $\hat{\mathbf{Y}}$ denote those from the full model and $\hat{\mathbf{Y}}^*$ those from the reduced model, it follows from (5.114a) that

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \quad (8.13)$$

and

$$\hat{\mathbf{Y}}_1^* = \mathbf{H}_1 \mathbf{Y}_1 \quad (8.14)$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ and $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T$.

Hence,

$$\begin{aligned} E(\hat{\mathbf{Y}}_1^*) &= \mathbf{H}_1 E(\mathbf{Y}) \\ &= \mathbf{H}_1 (\mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2) \\ &= \mathbf{H}_1 \mathbf{X}_1 \beta_1 + \mathbf{H}_1 \mathbf{X}_2 \beta_2 \\ &= \mathbf{X}_1 \beta_1 + \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2 \beta_2 \\ &= \mathbf{X}_1 \beta_1 + \mathbf{X}_1 \mathbf{A} \beta_2 \end{aligned}$$

So, $\hat{\mathbf{Y}}_1$ is generally a biased estimate of \mathbf{Y} unless the regression variables are orthogonal.

Similar result exists for the variances of $\hat{\mathbf{Y}}_1^*$ as for $\hat{\beta}_1^*$.

8.3 Criteria Functions

Before describing various algorithms for choosing the best subset (or best subsets) we examine a number of criteria for evaluating competing models.

Here we introduce some of the useful ones which are frequently used in selection procedures. Basically there are two different types of criteria functions depending on their usage: One is called a *selection function* which is a statistic which can be used to choose a specific model. The second type of criteria function is called an *assessment function*, which is used for assessing the performance of the model in terms of its intended use.

8.3.1 Coefficient of Multiple Determination R^2

Assume that the full model contains T variables (T for total) including the intercept and we have a subset of p variables including the intercept. Then, as in Chapter 5 we can define R_p^2 , the R^2 value for the p -variable model by

$$R_p^2 = \frac{SSR_p}{SST} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SSE_p}{SST} \quad (8.15)$$

where SSR_p and SSE_p denote the regression sum of squares and the residual sum of squares, respectively. It should be noted that since the corrected total sum of squares, SST , is constant for all regression models, R_p^2 cannot decrease as additional independent variables are introduced into the regression model. Thus the maximum value of R_p^2 will be attained when all possible variables are included in the model. With this in mind, our goal in utilizing R_p^2 is to compare alternative models so that we may determine when the introduction of additional regressor variables does not contribute a substantial increase in R_p^2 . To use R_p^2 as an assessment criterion we note the following properties of R_p^2 .

Properties of R_p^2

- (1) Increasing the number of regressors will increase the coefficient of determination R_p^2 .
- (2) For a given *residual mean square*, $s^2 = SSE_p/(n-p)$, denoted by RMS_p , the magnitude of R_p^2 depends on the magnitude of the regression coefficients. That is, R_p^2 is not scale invariant.

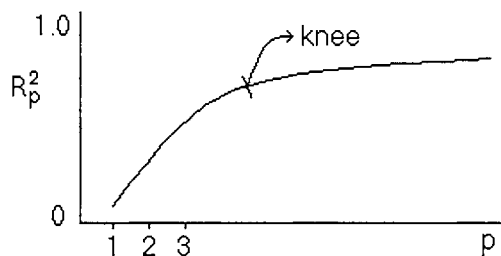
As we explained above, since we use the same set of data for different regression models SST is the same for each regression. This means that the R^2 statistic is not an absolute measure but a relative measure of goodness-of-fit. Furthermore, introducing an additional regressor increases R^2 , so it is not just a matter of finding the subset with maximum R^2 but rather that of finding a suitable subset with a high R^2 . To overcome some of these difficulties a modification of R^2 , the *adjusted coefficient of determination* \bar{R}_p^2 , defined by in Chapter 5

$$\begin{aligned} \bar{R}_p^2 &= 1 - \frac{(n-1)(1-R_p^2)}{n-p} \\ &= 1 - \frac{RMS_p}{SST/(n-1)} \end{aligned} \quad (8.16)$$

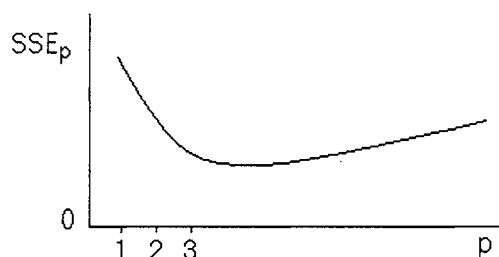
is often used instead of R_p^2 .

From (8.16), we see that maximizing \bar{R}_p^2 is equivalent to minimizing the residual mean squares (RMS_p). A criticism of both R_p^2 and \bar{R}_p^2 is that neither incorporates into the decision/selection criterion any consideration of the effects of incorrect model specification.

If a large number of subsets are examined, then a graphical technique can be used to examine R_p^2 . Plotting R_p^2 versus p generally results in a curve of the form in Figure 8.1. Values of R_p^2 near the “knee” of the curve suggest that models with those number of variables are good models.

Figure 8.1: Plot of R_p^2 versus p

Similarly, a plot of SSE_p versus p generally looks like Figure 8.2. Values of SSE_p near the bottom of the curve suggest good models as measured by R_p^2 or SSE_p .

Figure 8.2: Plot of SSE_p versus p

8.3.2 Mallows' C_p

Since R_p^2 essentially measures only the effective of bias, it is perhaps somewhat better to have a criterion which takes into account both precision and bias. Since a subset model is biased, it is appropriate to look for models which minimize $E(SSE_p)$, the expected mean square error.

Let \hat{Y}_p be the vector of predicted values from the p -variable model. If \mathbf{Y} is the vector of observations, then

$$\begin{aligned}
 MSE_p &= E(SSE_p) \\
 &= E\left[\left\langle \mathbf{Y} - \hat{\mathbf{Y}}_p, \mathbf{Y} - \hat{\mathbf{Y}}_p \right\rangle\right] \\
 &= E\left[\left(\mathbf{Y} - \hat{\mathbf{Y}}_p\right)^T \left(\mathbf{Y} - \hat{\mathbf{Y}}_p\right)\right].
 \end{aligned} \tag{8.17}$$

Recalling from (5.114a) that $\hat{\mathbf{Y}}_p = \mathbf{H}_p \mathbf{Y}$ where \mathbf{H}_p is the hat matrix for the p -variable model, then,

$$\left(\mathbf{Y} - \hat{\mathbf{Y}}_p\right)^T \left(\mathbf{Y} - \hat{\mathbf{Y}}_p\right) = \mathbf{Y}^T (\mathbf{I}_n - \mathbf{H}_p) \mathbf{Y}. \tag{8.18}$$

Hence,

$$MSE_p = E [\mathbf{Y}^T (\mathbf{I}_n - \mathbf{H}_p) \mathbf{Y}]. \quad (8.19)$$

Since this is quadratic form in \mathbf{Y} we use Theorem 4.14 to get (i.e., $\mathbf{A} = \mathbf{I}_n - \mathbf{H}_p$)

$$MSE_p = \langle E(\mathbf{Y}), (\mathbf{I}_n - \mathbf{H}_p) E(\mathbf{Y}) \rangle + \text{tr}[(\mathbf{I}_n - \mathbf{H}_p) \Sigma(\mathbf{Y}) (\mathbf{I}_n - \mathbf{H}_p)]. \quad (8.20)$$

Since $E(\mathbf{Y}) = \mathbf{X}\beta$, $\Sigma(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, and $(\mathbf{I}_n - \mathbf{H}_p)^2 = (\mathbf{I}_n - \mathbf{H}_p)$, (8.20) becomes

$$MSE_p = \langle \mathbf{X}\beta, (\mathbf{I}_n - \mathbf{H}_p) \mathbf{X}\beta \rangle + \sigma^2 \text{tr}(\mathbf{I}_n - \mathbf{H}_p). \quad (8.21)$$

Because \mathbf{H}_p has rank p , $\text{tr}(\mathbf{I}_n - \mathbf{H}_p) = n - p$ so that

$$\begin{aligned} MSE_p &= \langle \mathbf{X}\beta, (\mathbf{I}_n - \mathbf{H}_p) \mathbf{X}\beta \rangle + \sigma^2 (n - p) \\ &= SSE_B + \sigma^2 (n - p) \end{aligned} \quad (8.22)$$

where SSE_B is the bias term, which is zero if $p = T$. From this, a reasonable assessment criterion would be to choose p -variable models which minimize $E(SSE_p)$.

Unfortunately, since this depends on the unknown parameter vector β , this is not immediately useable. To overcome this difficulty Mallows (1964, 1966, 1973) considered the closely related statistic

$$C_p = \frac{SSE_p}{s^2} + 2p - n \quad (8.23)$$

where s^2 is the MSE from the full model with T variables. From (8.23) we have

$$E(C_p) = E\left(\frac{SSE_p}{s^2}\right) + 2p - n. \quad (8.24)$$

From above, a model with small bias has $E(SSE_p) \simeq (n - p) \sigma^2$ and assuming $s^2 \simeq \sigma^2$ then

$$E(C_p) \simeq \frac{\sigma^2 (n - p)}{\sigma^2} + 2p - n = p. \quad (8.25)$$

From this discussion assuming that models with small bias are desirable, it is reasonable to select models with "small" C_p . In fact models which minimize C_p such that $C_p \simeq p$.

Again, if there is a large number of variables it is convenient to plot C_p versus p superimposed over the line $C_p = p$ as shown in Figure 8.3.

Points close to the line indicate good models as measured by C_p . An additional form for C_p is given by

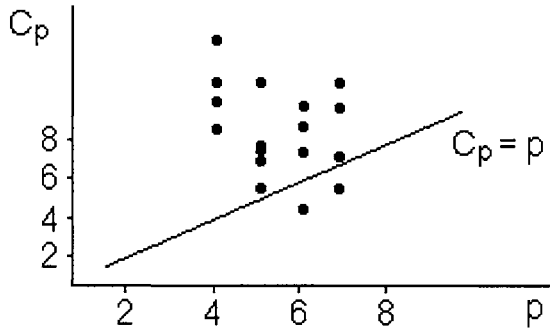
$$C_p = \frac{SSE_p - SSE_T}{s^2} + 2p - T. \quad (8.26)$$

In fact,

$$\frac{SSE_p - SSE_T}{s^2} = \frac{SSE_p - SSE_T}{SSE_T / (n - T)} \quad (8.27)$$

so that (8.27) becomes

$$\frac{SSE_p - SSE_T}{SSE_T / (n - T)} - (n - T) + 2p - T = \frac{SSE_p - SSE_T}{s^2} + 2p - n = C_p. \quad (8.28)$$

Figure 8.3: Plot of C_p versus p

Since $(SSE_p - SSE_T)/s^2 = F_p$ is the F -statistic for testing the significance of the p -variable model, then it follows from (8.28) that

$$C_p = (T - p)(F_p - 1) + p. \quad (8.29)$$

Again, for the full model it follows from (8.29) that $C_T = T$.

It may be, however, that the full model is poorly specified and that the resulting mean square error is inflated. In such cases, the value of C_p can be negative. This is not to say that the model with the smallest C_p is poor; it merely states that the full model is poorly specified.

Properties of C_p

- (1) C_p depends only on the usual regression calculations, namely SSE_p , s^2 , p , and n . This is the basis for the use of C_p in fast all possible regressions calculations.
- (2) C_p measures the difference in fitting errors between the full and subset models.
- (3) C_p consists of two parts: a random part F_p and a penalty p . This means that there is a trade-off between decreasing F_p and adding variables. (Due to this the C_p criterion is sometimes referred to as a *penalized method* of choosing a model.)
- (4) C_p is closely related to the adjusted coefficient of determination \bar{R}_p^2 [71]. Since $1 - \bar{R}_p^2 = \frac{n}{n-p}(1 - R_p^2)$, and estimating σ^2 by $\hat{\sigma}^2 = \frac{SSE_T}{n-T}$, we have

$$C_p = \frac{(n-T)SSE_p}{SSE_T} + 2p - n \quad (8.30)$$

or

$$1 + \frac{C_p - p}{n - p} = \frac{(n-T)SSE_p}{(n-p)SSE_T} = \frac{1 - \bar{R}_p^2}{1 - \bar{R}_T^2},$$

where \bar{R}_T^2 is the adjusted coefficient of determination for a model containing all T parameters.

One disadvantage of C_p is that it seems to be necessary to calculate C_p for all, or most, of the possible subsets, to allow interpretation. For other illustrations and comments and further examples of the use of C_p , the reader is referred to Gorman and Toman [44], Mallows [82], or Daniel and Wood [22].

8.3.3 The PRESS Statistic

So far we have discussed selection criteria based on assessing the fit. However, if prediction is an important consideration we might consider selection criteria which assesses the predictive properties of the model. For example, as we shall see in Example 8.1 assessment criteria such as R_p^2 , \bar{R}_p^2 or C_p can often provide a number of different models with similar values of these criteria. For further selection one can consider how well each of these models predicts at new points and then one can further narrow our selection based on this property. However, as noted previously in Chapter 6, generally we will not know the true value of the model at the new data points, but this can be overcome to some extent by using cross-validation. That is, we omit one (sometimes more) data point, refit and then predict using the model with the i -th observation deleted to predict the value at this point. As in Chapter 6, let $\hat{y}_{(-i)}$ be this predicted value and then the i -th PRESS residual is given by $\hat{\varepsilon}_{(-i)} = y_i - \hat{y}_{(-i)}$. In [2] Allen and Stone in [111] proposed using the PRESS statistic

$$PRESS = \sum_{i=1}^n \hat{\varepsilon}_{(-i)}^2 \quad (8.31)$$

as a measure of the predictive power of the model.

As noted in Section 6.4

$$\hat{\varepsilon}_{(-i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}} \quad (8.32)$$

where $\hat{\varepsilon}_i$ represents the i -th residual from the full model and h_{ii} is the i -th diagonal element of the hat matrix \mathbf{H} . Hence,

$$PRESS = \sum_{i=1}^n \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2 \quad (8.33)$$

which can be computed without having to refit. Letting $PRESS_p$ be this value for a p -variable model we have

$$PRESS_p = \sum_{i=1}^n \left(\frac{\hat{\varepsilon}_{i,p}}{1 - h_{ii,p}} \right)^2 \quad (8.34)$$

where $\hat{\varepsilon}_{i,p}$ is the i -th residual from the p -variable model and $h_{ii,p}$ is the i -th diagonal element of

$$\mathbf{H}_p = \mathbf{X}_p (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T. \quad (8.35)$$

Letting $\mathbf{D}_p = \text{diag}[\mathbf{I}_n - \mathbf{H}_p]$, $PRESS_p$ can be written as

$$PRESS_p = \langle \hat{\varepsilon}_p, \mathbf{D}_p^{-2} \hat{\varepsilon}_p \rangle = \hat{\varepsilon}_p^T \mathbf{D}_p^{-2} \hat{\varepsilon}_p \quad (8.36)$$

where $\hat{\varepsilon}_p = (\mathbf{Y} - \hat{\mathbf{Y}}_p)^T$ is the residual vector. It is suggested that the best predictive model is one that minimizes $PRESS_p$.

8.3.4 Standardized Residual Sum of Squares

In [102] Schmidt suggested a criterion function for assessment that is closely related to $PRESS_p$, which is called the *standardized residual sum of squares* (RSS^*), defined by

$$RSS_p^* = \hat{\mathbf{e}}_p^T \mathbf{D}_p^{-1} \hat{\mathbf{e}}_p \quad (8.37)$$

where $\hat{\mathbf{e}}_p$ and \mathbf{D}_p are the same as in equation (8.36).

Intuitively the true model minimizes $E(RSS_p^*)$. Inspection of (8.36) and (8.37) shows that they are both weighted sums of squares of the residuals and, hence, direct comparison of these two functions with those previously discussed is difficult. However, it would appear that for large samples both (8.36) and (8.37) would be close to RSS_p because one can expect that \hat{y}_i would be approximately equal to the estimator $\hat{y}_{(-i)}$.

8.3.5 Other Criteria

Besides the above criteria functions, many other criteria have been suggested for selection procedures. We describe some of them briefly.

(1) Tukey's rule

The rule says to choose the set of regressors which yields a minimum of s^2/ν , where s^2 is the mean square error and $\nu = n - p$ is the degrees of freedom of the model chosen. See Anscombe and Tucky [4] for more details.

(2) Average Estimated Variance

The *average estimated variance* (AEV) criterion was suggested by Helms [53]. The key idea is to average the prediction variance over the whole regression region of interest, rather than just for the data points given, using a weight function which attaches more weight to the more important points in the region. In the special case when the moment matrix, $\mathbf{M} = (\mathbf{X}^T \mathbf{X})/n$, Helms ([53], p. 265) explored the monotonic relationship between

$$AEV = \frac{p \cdot RSS_p}{n(n-p)} \quad (8.38)$$

and C_p for subsets with p variables. Helms ([53], p. 269) also questioned the practice of always including an intercept term β_0 in the model by saying that "our experience has indicated intercept terms are frequently primary contributors to variance but their absence often leads to only small contributions to bias."

In summary, although many selection criteria have been developed, there is no general theory as to which one(s) to use in a given situation. For further information we recommend seeing Hocking [57], Seber [104] and Thompson [113, 114]. Further approaches are given below.

8.4 Various Methods for Model Selection

In this section we discuss various computational techniques for model selection. In [57] Hocking pointed out that three distinct ingredients of techniques for model selection procedures in multiple regression analysis can be identified:

- (a) the computational techniques used to identify a set of possible models to be considered;
- (b) the criterion function used to select a particular model;
- (c) the estimation of parameters in the chosen model.

Usually all three of these are put together but we shall discuss them separately. On estimation, several authors (Rencher and Pun [98], Copas [21]) have pointed out that if least squares estimation is used in any situation where the same data is used for selecting the model as for estimating the parameters, and there is competition between models, then the least squares estimators are biased.

According to Miller [86] and Berk [7] this bias can be substantial. Some comments on how to cope with the bias are made later. For the rest of this section we consider computational techniques of the stepwise type, and the results of these are sets of possible models to which we need to apply a selection criterion.

8.4.1 Evaluating All Possible Regressions

In selecting the best regression model, the first approach we consider is to evaluate all possible linear regression models. The procedure requires the fitting of every possible regression model which includes the intercept β_0 and any number of the regressor variables $x_i, i = 1, 2, \dots, T$. Therefore there are 2^T total potential models to be considered. For the assessment of all possible linear regression models we shall focus upon the following criteria.

- (a) the R_p^2 statistic from the least squares fit;
- (b) s^2 , the residual mean square, and
- (c) the C_p statistic.

Example 8.1 (Hald Cement data) Let us reconsider the Hald cement data ($n = 13$) that was used in Example 5.16. We will use these data to illustrate the “all possible regressions” approach to variable selection. Since the number of variables under consideration is four, there will be a total of $2^4 = 16$ possible regression models that include the term β_0 . The value of m indicates the number of regressors in the model ($p = m + 1$). The results of fitting these 16 models are shown in Table 8.1.

We first evaluate the subset models using the R_p^2 criterion from Table 8.1 and Figure 8.4. We observe that there is great similarity among the R_p^2 values for the various regression models. For the two-parameter model ($p = 2; \beta_0$ and one regressor) the best fit clearly occurs when x_4 is included ($R_p^2 = 0.6746$) and next is when x_2 is included ($R_p^2 = 0.6663$). For three-parameter models, x_1, x_2 and x_1, x_4 show relatively higher R_p^2 values. In addition, for the four-parameter models, x_1, x_2, x_3 and x_1, x_2, x_4 show higher ones. Finally, with all five parameters considered, $R_5^2 = 0.9824$. From Figure 8.5 we can see that R_p^2 steadily rises until three parameters are included in the model and then does not exhibit any substantial increase as additional parameters are added. So as we see, the best two-parameter model (x_1, x_2), has 0.9744 while the best three-parameter model (x_1, x_2, x_4) has 0.9765. Therefore, in the interest of simplicity, the researcher

would decide to select the regression model which contains only x_1, x_2 since inclusion of x_4 leads to a negligible increase in R_p^2 .

Table 8.1 Summary of All Possible Regressions for Hald data

Model No.	Regressors in model	m	SSE_p	R_p^2	\overline{R}_p^2	$MSE_p (= s^2)$	C_p
1	None (β_0)	0	2715.7635	0	0	226.3136	442.92
2	x_1	1	1265.6867	0.5340	0.4916	115.0624	202.55
3	x_2	1	906.3363	0.6663	0.6359	82.3942	142.49
4	x_3	1	1939.4005	0.2859	0.2210	176.3092	315.16
5	x_4	1	883.8669	0.6746	0.6450	80.3515	138.73
6	x_1, x_2	2	57.9045	0.9787	0.9744	5.7904	2.68
7	x_1, x_3	2	1227.0721	0.5482	0.4578	122.7073	198.10
8	x_1, x_4	2	74.7621	0.9725	0.9670	7.4762	5.50
9	x_2, x_3	2	415.4427	0.8470	0.8164	41.5443	62.44
10	x_2, x_4	2	868.8801	0.6801	0.6161	86.8880	138.23
11	x_3, x_4	2	175.7380	0.9353	0.9224	17.5738	22.37
12	x_1, x_2, x_3	3	48.1106	0.9823	0.9764	5.3456	3.04
13	x_1, x_2, x_4	3	47.9727	0.9823	0.9765	5.3303	3.02
14	x_1, x_3, x_4	3	50.8361	0.9813	0.9750	5.6485	3.50
15	x_2, x_3, x_4	3	73.8145	0.9728	0.9638	8.2017	7.34
16	x_1, x_2, x_3, x_4	4	47.8636	0.9824	0.9736	5.9829	5.00

Also, we calculate the simple correlations among the four regressor variables and the response.

Table 8.2 Correlation Matrix for Hald’s data

	y	x_1	x_2	x_3	x_4
y	1.0				
x_1	0.731	1.0			
x_2	0.816	0.299	1.0		
x_3	−0.535	−0.824	−0.139	1.0	
x_4	−0.821	−0.245	−0.973	0.030	1.0

However, examining all possible regressions easily becomes a laborious and computationally burdensome task if a large number of regressors are under consideration. For example, when the number of the independent variables $T = 10$, there will be 1,024 possible models to consider! Thus, without examining all possible models, other search procedures have been developed in order to find the “best” subset of variables by adding or removing variables one at a time. These methods are generally called stepwise procedures, which can be categorized according to the method of adding or removing variables: (1) *backward elimination*, (2) *forward selection*, and (3) *stepwise regression* which is a combinational of procedures (1) and (2). We now describe these in detail.

8.4.2 Backward Elimination

The backward elimination procedure starts with the full model and removes one variable at a time without adding variables. One includes all T possible regressor variables, and attempts to eliminate them from the model one at a time until no removal occurs. Since

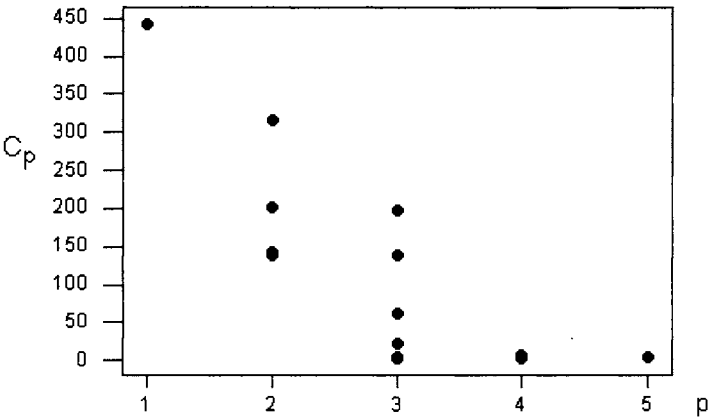


Figure 8.4: Plot of C_p versus p for Hald data

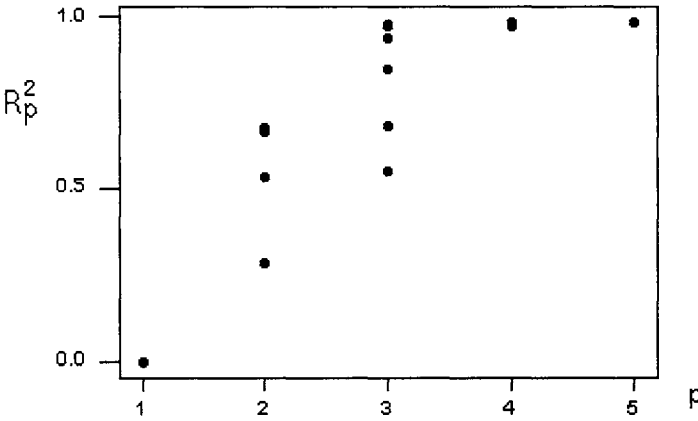


Figure 8.5: Plot of R_p^2 versus p for Hald data

the backward elimination method seeks only to remove variables from the model, the variable with the smallest incremental contribution to the regression is tested at each step to determine whether it can be eliminated from the model.

This procedure seems to be more economical because it evaluates fewer numbers of models than the previous approach. Even though the models examined are more complex and require more computing time, often the backward elimination procedure is regarded as a good variable selection procedure. We now summarize the basic steps in the method as:

- (1) A regression model with all variables included is computed.
- (2) The partial F -statistic (or t statistic) is calculated for each regressor as if it just entered into the model. (This is to evaluate the incremental contribution of the regressor).
- (3) If any of these partial F -statistics (or t statistics) is smaller than the critical value F_{out} which is a prescribed level of significance (i.e., F -to-remove), remove the variable from the model and the partial F -value for this new model with $p - 1$ variables is recalculated.
- (4) Repeat (2) and (3) and the procedure will be terminated when the smallest partial F -value is not smaller than the critical value F_{out} .

Example 8.2 (Hald data) We use the Hald cement data to illustrate the backward elimination procedure. Before running the computer package (MINITAB and many other statistical packages have these functions), we take $\alpha = 0.05$ for the cutoff point, so $F_{\text{out}} = F_{1,n-T,0.05}$ (or $t_{\text{out}} = t_{n-T,0.05}$). Then any regressor will be dropped from the model if its partial F -statistic is less than F_{out} . The backward elimination process starts with the full model and calculates the partial F -values (or equivalently t values) for all individual regressors. We now write out the detailed description of the procedure.

First, the regression model containing all four variables is fit. From the results, this is

$$\hat{y} = 62.4 + 1.55x_1 + 0.510x_2 + 0.102x_3 + 0.144x_4$$

with the overall F -value 111.48. Denoting the partial F -value $F(x_j|x_i)$ by $F_{j|i}$, the four F -values:

$$\begin{aligned} F_{1|2,3,4} &= 4.326, & F_{2|1,3,4} &= 0.49, \\ F_{3|1,2,4} &= 0.0196 & F_{4|1,2,3} &= 0.04. \end{aligned}$$

Recall that the extra sum of squares to obtain all these F -values is given by (5.247). For instance, $s = 2.446$ so that $F_{1|2,3,4} = [(73.815 - 47.863)/1]/5.983 = 4.326$. We also note that one can get these values of F by squaring the t -values in the regression of Y onto x_1, x_2, x_3 and x_4 , that is, $F_{1,v} = t_v^2$ (v is the degrees of freedom). Since $F_{\text{out}} = F_{1,8,0.05} = 5.32$ from Table A.4b, we drop the variable X_3 which has the smallest partial F from the model.

Next, we consider the reduced model with three variables, x_1, x_2 and x_4 . The result shows that the fit is

$$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$$

and $s = 2.309$ with the overall F -value 166.83 which is significant at 1% ($F_{3,9,0.01} = 13.90$). Since the three partial F -values are

$$F_{1|2,4} = 154.01, \quad F_{2|1,4} = 5.02, \quad F_{4|1,2} = 1.88,$$

by comparing to the critical value $F_{1,9,0.05} = 5.12$, we eliminate the variable x_4 .

We now work on with the model $\hat{Y} = f(x_1, x_2)$. One can find

$$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2,$$

which provides $s = 2.406$ and the overall F -statistic is $229.50 > F_{2,10,0.01} = 14.91$. We find that both partial F -values for x_1, x_2 exceed the critical value $F_{1,10,0.05} = 4.96$. Therefore, the backward elimination selection procedure terminates, yielding the final model

$$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2.$$

The following Table 8.2 summarizes the steps in the backward elimination procedure based on the output from MINITAB.

Table 8.2 Backward Elimination Method for the Hald Data

Step	Constant	x_1	x_2	x_3	x_4	R_p^2, C_p	F_{out} & decision
1	62.41 (t -value)	1.55 2.08	0.510 0.70	0.10 0.14	-0.14 -0.20	0.982 5.0	$F_{1,8,0.05} = 5.32$ $\Rightarrow x_3$ removed
2	71.65 (t -value)	1.45 12.41	0.416 2.24	- -	-0.24 -1.37	0.982 3.0	$F_{1,9,0.05} = 5.12$ $\Rightarrow x_4$ removed
3	52.58 (t -value)	1.47 12.10	0.662 14.44	- -	- -	0.979 2.7	$F_{1,10,0.05} = 4.96$ \Rightarrow Stopped

8.4.3 Forward Selection

With this approach, we start with no variables in the model (other than the intercept), and add one variable at a time which gives the largest increase in the regression sum of squares. This procedure does not permit the removal of a variable from the model once it has been entered. The method is summarized as follows:

- (1) Calculate the individual sample correlations or F -statistics for testing the significance of adding one variable to the regression.
- (2) If any of these partial F -statistics exceeds the critical value F_{in} which is a prescribed level of significance (i.e., F -to-enter), enter the variable with the largest F -value into the model.
- (3) Repeat (2). That is, the regressor having the largest partial F -value (or equivalently the highest partial correlation with y given the other regressors already in the model) is added to the model if its F -value exceeds the F_{in} value.
- (4) The selection procedure will terminate if either the partial F -statistic at a particular step does not exceed F_{in} or if the last candidate variable is added to the model.

Although this simplifies the model selection procedure, oftentimes it unfortunately leads to the inclusion of variables that do not make a significant contribution once other independent variables are entered in the regression model.

Example 8.3 (Hald data) We will conduct the forward selection procedure using the Hald cement data. Let's take the value of $\alpha = 0.10$. Then $F_{in} = F_{1,n-p,\alpha}$ is the critical value in each step in the procedure.

As a first step, we chose x_4 because it has the largest absolute simple correlation ($r_{y,4} = -0.821$) with the response variable y , and the model using x_4 gives the F -value $= 22.80 > F_{1,11,0.10} = 3.296$. Then, we continue to proceed according to the steps described in the above until no other variable exceeds the critical value.

We summarize the steps in the forward selection procedure in Table 8.3 based on the output from MINITAB.

Table 8.3 Forward Selection Method for Hald Data

Step	Constant	x_1	x_2	x_3	x_4	R_p^2, C_p	F_{in} & decision
1	117.57 (t -value)	- -	- -	- -	-0.738 -4.77	0.675 138.7	$F_{1,11,0.10} = 3.23$ $\Rightarrow x_4$ entered
2	103.10 (t -value)	1.44 10.40	- -	- -	-0.614 -12.62	0.972 5.5	$F_{1,10,0.10} = 3.29$ $\Rightarrow x_1$ entered
3	71.65 (t -value)	1.45 12.41	0.42 2.24	- -	-0.237 -1.37	0.982 3.0	$F_{1,9,0.10} = 3.36$ $\Rightarrow x_2$ entered
							Stopped

Hence, we conclude that the final model using forward selection procedure contains x_1, x_2 and x_4 , which is

$$\hat{y} = 71.65 + 1.45x_1 + 0.42x_2 - 0.237x_4.$$

We also note that this may be different from the model we found using backward elimination selection procedure even if we use the same $\alpha = 0.05$. That is, the same α does not guarantee the same results among different selection procedures. In fact, if we would have used $\alpha = 0.05$, the forward selection procedure leads to a model that contains the variables x_1 and x_2 .

8.4.4 The Stepwise Regression Procedure

Perhaps this is the most widely used procedure for model selection. This procedure employs a series of tests (t or F) to check for the significance of the regressors entered into, or removed from, the model. Since the stepwise regression procedure is a combination of forward selection and backward elimination, the procedure requires two cutoff values, F_{in} for inclusion and F_{out} for removal.

In each step, all regressors entered into the model previously are reassessed via their partial F -statistics. A regressor added at an earlier step can now be eliminated if its partial F -statistic is smaller than F_{out} . Likewise, a regressor once dropped before may be added again into the model if its recalculated partial F -statistic is larger than F_{in} . Often it is convenient to choose $F_{in} = F_{out}$. If we take $F_{in} > F_{out}$, then it becomes more difficult to add a variable than to remove one. Note that when only one variable is being considered, that $(t\text{-ratio})^2 = F\text{-ratio}$ and thus the t -test and F -test are equivalent. The following is a description of the basic algorithm.

- (1) The procedure begins with the variable chosen first, say $x_{(1)}$, that is most highly correlated with the response variable Y . This variable is also the one that produces the largest partial F -value. If the F -statistic for this model is larger than F_{in} , the model includes the variable. Otherwise, the process terminates with no regressors included in the model.
- (2) Regress Y on $x_{(1)}$, and a partial F -test is computed for each of the $p - 1$ remaining variables given $x_{(1)}$. If the largest partial F -value $> F_{\text{in}}$, then the second variable, say $x_{(2)}$, would be included. Otherwise, the process is terminated, and only $x_{(1)}$ is included in the model.
- (3) We now determine whether any of the variables already included are no longer important, given that others have subsequently been added. If the partial F -value $> F_{\text{in}}$, keep it in the model. If the partial F -value $< F_{\text{out}}$, drop it from the model.
- (4) Repeat (3) until no other variables are to be entered or removed.

Example 8.4 (Hald data) We will illustrate stepwise regression procedure using the Hald cement data. We take the size of $\alpha = 0.10$ for both F_{in} and F_{out} . At the beginning stage, consider $F_{\text{in}} = F_{1,11,0.10} = 3.23$. The first choice will be to include x_4 because $r_{y,4} = -0.821$ is the highest from the previous result - forward selection. It is also the variable with the largest F -value $22.80 > F_{\text{in}} = F_{1,11,0.10} = 3.23$. Hence, the variable x_4 enters the model. The model is now

$$\hat{y} = 117.568 - 0.7382x_4,$$

and the overall F -statistic is 22.80 with $R^2 = 0.675$. At the second step, we calculate the three F -values: $F_{1|4} = 108.22$, $F_{2|4} = 0.17$, and $F_{3|4} = 40.29$. Since $F_{1|4} > F_{\text{in}} = F_{1,10,0.10} = 3.29$, x_1 is added to the model. Then, for possible removal of the variables, recall that $F_{\text{out}} = F_{1,10,0.10} = 3.29$ at this stage. If the partial F -value for previously entered variables (here we have x_4) is smaller than F_{out} , then the variable would be removed. The two partial F -values we need to consider are: $F_{1|4} = 108.22$ and $F_{4|1} = 159.30$. But both are larger than $F_{\text{in}} = F_{1,10,0.10} = 3.29$, so we retain both x_1 and x_4 . Therefore, the model becomes

$$\hat{y} = 103.10 + 1.440x_1 - 0.614x_4,$$

where $R^2 = 0.972$ and the overall F -statistic is 176.63, which is clearly significant at 1%.

The stepwise regression method now considers the next variable to add. Since $F_{2|1,4} = 5.03 > F_{3|1,4} = 4.24$, x_2 will be a candidate. Hence, the model we are considering is the one with variables (x_1, x_2, x_4) . The model has an overall F -statistic of 166.83 and $R^2 = 0.982$. Then, we need to check for possible deletion among the three partial F -values;

$$F_{1|2,4} = 154.01, \quad F_{2|1,4} = 5.02, \quad F_{4|1,2} = 1.88.$$

As we see, $F_{4|1,2} = 1.88 < F_{\text{out}} = F_{1,9,0.10} = 3.36$, so we delete variable x_4 . We note that x_2 cannot be eliminated because in order to move we must recompute the model $\hat{y} = f(x_1, x_2)$. Thus, since no other variables are to be entered or removed, the stepwise regression procedure terminates with a final model that contains the variables (x_1, x_2) which is

$$\hat{y} = 52.5773 + 1.4683x_1 + 0.6623x_2.$$

We present a summary in the following Table 8.4 of the steps in stepwise regression procedure based on the output from MINITAB.

Table 8.2 Stepwise Regression Method for Hald Data

Step	Constant	x_1	x_2	x_3	x_4	R_p^2, C_p	$F_{in}-F_{out}$ & decision
1	117.57 (t -value)	- -	- -	- -	-0.738 -4.77	0.675 138.7	$F_{in} = F_{1,11,0.10} = 3.23$ $\Rightarrow x_4$ entered
2	103.10 (t -value)	1.44 10.40	- -	- -	-0.614 -12.62	0.973 5.5	$F_{in} = F_{1,10,0.10} = 3.29$ $\Rightarrow x_1$ entered
3	71.65 (t -value)	1.45 12.41	0.416 2.24	- -	-0.237 -1.37	0.982 3.0	$F_{in} = F_{1,9,0.10} = 3.36$ $\Rightarrow x_2$ entered
4	52.58 (t -value)	1.47 12.10	0.662 14.44	- -	- -	0.979 2.7	$F_{out} = F_{1,9,0.10} = 3.36$ $\Rightarrow x_4$ removed
							Stopped

8.4.5 Selection of Models - An Overview

Criticism of Selection Procedures

We have examined several approaches to selecting the best subset in a regression model. It is important to note that none of the stepwise procedures discussed above can claim any kind of optimality, and these approaches do not necessarily result in the same set of variables in the model. Thus the researcher must be aware of the fact that there is often no “uniquely superior” or “best” regression model for a set of T regressor variables. Nevertheless, since all the stepwise procedures terminate with one final model, inexperienced analysts or novices may conclude that they have found a model that is optimal in some sense or that they must accept the model uncritically. The aim of selection is always to maximize the ability to find out all of the “relevant” information that is hidden in the data. Therefore, several approaches can be utilized or combined in appropriate ways in attempting to find a “best” regression model, such as two-stage selection procedures.

Further criticisms can be found in Gorman and Toman [44], Mantel [83], and Hocking [57].

Choice of Stopping Rules

For the choice of stopping rules - F_{in} or F_{out} - it should be recognized that in the stepwise procedures the choice of a small α will limit the number of models that would be explored. To avoid an early termination of the selection procedure, it is better to choose α larger than typical values 0.01 or 0.05, say 0.20 or 0.25 (but bear in mind that the size of the Type I error becomes larger!). Similarly, a much larger value of α , could also be used for F_{out} in order to increase the number of models explored in the algorithm.

Further Comments in Selection Procedures

If the number of variables T is not too large, say about a dozen, either all possible regressions or backward elimination may be preferable in terms of computation. Forward selection can be seriously misleading because of the restriction of adding one variable at a time. Also it may be using badly inflated estimates of σ^2 , whereas a watch can be kept on this with backward elimination.

If multicollinearity is known to be present then ridge regression, or some other biased estimation procedure, can be used as a method of selecting variables. (See the next chapter.)

Consequently, the selection procedure leads to two separate testing problems: one is the problem of whether the estimated regression coefficients are significantly different from zero and the other is the problem of testing the difference between models. Forsythe, Engelman, Jennrich and May [35] proposed a permutation test for the first problem, and the second problem is a multiple comparison problem of the Scheffé type [101]. See Spjøtvoll [107, 108] for more details. Unfortunately both authors do not deal with selection bias in the estimators.

Once a best model has been found, a thorough residual analysis should be performed to evaluate the aptness of the model. In conclusion, although the various calculation approaches provide guidelines for variable selection, the model ultimately developed should take into consideration such factors as its simplicity, interpretability/predictive ability, and the usefulness of the variables.

8.5 Exercises

- 8.1** Gorman and Toman [44] discussed an experiment in which the rate of rutting was measured on thirty-one experimental asphalt pavements. Five predictor variables were used to specify the conditions under which each asphalt was prepared, while a sixth dummy variable was used to express the difference between the two separate blocks of runs into which the experiment was divided.

The multiple regression model used to fit the data was:

$$Y = \beta_0 + \sum_{j=1}^6 \beta_j x_j + \varepsilon \quad (8.39)$$

where

Y = log(change of rut depth in inches per million wheel passes),

x_1 = log(viscosity of asphalt),

x_2 = per cent asphalt in surface course,

x_3 = per cent asphalt in base course,

x_4 = dummy variable to separate two sets of runs,

x_5 = per cent fines in surface course, and

x_6 = per cent voids in surface course.

You may assume that Equation (8.39) is “complete” in the sense that it includes all the relevant terms. Your assignment is to select a suitable subset of these terms as the “best” regression equation in the circumstances.

Using Table 8.5, answer the following questions and show the steps and information that led to your answer.

- (a) What is R^2 for the model with variables 1 and 2 in it? Then, what is s^2 ?

- (b) Calculate the C_p statistic value for the model with variables 1, 2, 3, and 4 in it? Give a comment.
- (c) Select a suitable subset of independent variables as the “best” regression equation if the backward elimination procedure is used. Take $\alpha = 0.1$.

Table 8.5 Residual Sums of Squares (RSS) for All Possible Models

Model*	RSS	Model*	RSS	Model*	RSS	Model*	RSS
-	11.058	5	9.922	6	9.196	56	7.680
1	0.607	15	0.597	16	0.576	156	0.574
2	10.795	25	9.479	26	9.192	256	7.679
12	0.499	125	0.477	126	0.367	1256	0.364
3	10.663	35	9.891	36	8.848	356	7.678
13	0.600	135	0.582	136	0.567	1356	0.561
23	10.168	235	9.362	236	8.838	2356	7.675
123	0.498	1235	0.475	1236	0.365	12356	0.364
4	1.522	45	1.397	46	1.507	456	1.352
14	0.582	145	0.569	146	0.558	1456	0.553
24	1.218	245	1.030	246	1.192	2456	1.024
124	0.450	1245	0.413	1246	0.323	12456	0.313
34	1.453	345	1.383	346	1.437	3456	1.342
134	0.581	1345	0.561	1346	0.555	13456	0.545
234	1.041	2345	0.958	2346	0.995	23456	0.939
1234	0.441	12345	0.412	12346	0.311	123456	0.307

* Variable numbers included in the regression model with a β_0 term.

- 8.2** Using Table 8.5, explain the forward selection procedure in selecting a suitable subset of independent variables as the best regression model. Take $\alpha = 0.05$.
- 8.3** Using Table 8.5, select a suitable subset of independent variables as the “best” regression equation if the stepwise regression procedure is used. α values are 0.5 for entry (or in) and 0.1 for remove (or out).
[Hint: The residual for the regression containing no predictors but only β_0 will give you the corrected total sum of squares (TSS).]
- 8.4** (Hocking [56]) Show that $C_p \leq p$ if and only if $F \leq 1$ where F is the F -statistic for testing the hypothesis that $m + 1 - p$ of the regression coefficients β_j are zero.
- 8.5** Show that for $m = 1$, the diagonal term of the HAT matrix is

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

[Hint: The diagonals are measures of standardized squared distance.]

- 8.6** (Gorman and Toman (1966, p. 50) [44]) Suppose that we wish to omit the regressor x_j from a multiple regression model with p parameters. If F_j is the F -statistic for testing $H_0 : \beta_j = 0$, show that

$$C_{p-1} = \frac{RSS_p}{\hat{\sigma}^2 (n - p)} F_j + C_p - 2.$$

Chapter 9

Multicollinearity: Diagnosis and Remedies

9.1 Introduction

As we pointed out in Chapter 5 one of the more difficult problems that occurs in doing a regression analysis is that of dealing with multicollinearity. In Example 5.17 we illustrated some of the consequences of this phenomenon; i.e., the difficulty in interpreting the apparently contradictory facts of a good overall fit as indicated by a large value of R^2 and a significant observed F ratio, simultaneously with insignificant values for the individual t statistics, while in [76] it was shown how multicollinearity could seriously affect the accuracy of regression calculations. In addition, some of the problems with variable selection techniques as we observed in Chapter 8 are often associated with strong multicollinearity in the data.

In this Chapter we will examine this problem in greater detail. In particular, we will discuss methods for detecting multicollinearity, elaborate on its statistical consequences and examine some of the proposed remedies, particularly the method of *ridge regression*. This latter technique, which has spawned volumes of research in the past 25 years [28, 58, 59], is still controversial as are other forms of biased estimation [8, 116] and is by no means espoused by all statisticians as the work of Draper and Van Nostrand [28] and Draper and Smith [27] shows. However, because of its prominent role in current research and applications, no modern treatment of regression analysis would be complete without some discussion of its use. At present, it is fair to say that many of the computational problems associated with multicollinearity have been overcome through the development of more sophisticated computational techniques [5, 112] such as the QR decomposition and the advent of powerful computers. While the problem of building and interpreting models with this problem is far from being totally resolved [116] and perhaps never will be.

9.2 Detecting Multicollinearity

As we indicated in Chapter 5, multicollinearity occurs in a regression problem if the columns of the design matrix are “approximately” linearly independent. Clearly, this is a subjective notion, since the notion of “approximate” may vary from problem to problem and more importantly from analyst to analyst. Notwithstanding this vagueness, considerable effort has been devoted to at least partially quantifying this notion.

As pointed out in Belsley, Kuh and Welch [8] historically, many of the attempts to do this have been seriously flawed (some of these procedures will be discussed shortly). As they note, this problem is not unique to regression analysis; it is a classical problem in numerical analysis associated with solving any set of linear equations; that of ill-conditioning. In numerical analysis this problem has been discussed in great detail and they recommend the use of standard numerical techniques based on the SVD for diagnosing the problem. We shall, for the most part, follow their recommendations.

Before we begin our analysis it is important to decide which form the design matrix to use in defining and detecting multicollinearity since this issue has been a matter of some controversy itself. If there is an intercept in the model then five possibilities have been suggested;

- (i) Use the original design matrix

$$\mathbf{X}_1 = [\mathbf{1} \mid \mathbf{X}^*] \quad (9.1)$$

where \mathbf{X}^* is the $n \times m$ matrix of regressor variables and $\mathbf{1}$ is a column of ones.

- (ii) Use

$$\mathbf{X}_2 = [\mathbf{1} \mid \mathbf{X}_c^*] \quad (9.2)$$

where \mathbf{X}_c^* is the matrix of centered regressor variables.

- (iii) Use

$$\mathbf{X}_3 = [\mathbf{1} \mid \mathbf{X}_{sc}^*] \quad (9.3)$$

where \mathbf{X}_{sc}^* is the matrix of centered and scaled (as in Example 5.9) regressor variables.

- (iv) Use

$$\mathbf{X}_4 = [\mathbf{1}/\sqrt{n} \mid \mathbf{X}_s^*] \quad (9.4)$$

where \mathbf{X}_s^* is the matrix of scaled regressor variables obtained by dividing each element of the i -th column by its Euclidean length.

- (v) Use

$$\mathbf{X}_5 = \mathbf{X}_{sc}^* \quad (9.5)$$

BKW recommend using \mathbf{X}_3 because this enables one to detect approximate dependencies involving the intercept. On the other hand, Draper and Smith [27] and Montgomery and Peck [87] favor the use of \mathbf{X}_5 .

Since our focus in this Chapter will be more towards remedies rather than identifying specific collinearities we will perform all diagnostic procedures on \mathbf{X}_{sc}^* . This is then consistent with the recommended approach to *ridge regression* favored by Draper and

Van Nostrand [28]. Consequently, throughout the remainder of this chapter we will assume that the linear model is of the form

$$\mathbf{Y} = \mathbf{X}_{sc}^* \boldsymbol{\beta}_s + \boldsymbol{\varepsilon} \quad (9.6)$$

where \mathbf{Y} is also centered and scaled and the errors $\varepsilon_i, 1 \leq i \leq n$, are independent $N(0, \sigma^2)$. The transition between the least squares (and other) estimators of $\boldsymbol{\beta}$ in the original model $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}^* \boldsymbol{\beta}_s + \boldsymbol{\varepsilon}$ will be made in accordance with the prescription given in Example 5.9 when \mathbf{y} is also scaled and centered then $\hat{\boldsymbol{\beta}} = \mathbf{S}^2 \hat{\boldsymbol{\beta}}_s$ where \mathbf{S} is a scaling matrix. (See Example 5.9.) To avoid excessive notational complexity we will generally just use \mathbf{X} to refer to \mathbf{X}_{sc}^* and $\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}_s$.

Eigenvalue Criteria

Recall that in Chapter 5 we said that \mathbf{X} exhibits multicollinearity if there exists a vector $\mathbf{c} = (c_1, \dots, c_n)^T$ which has length one ($\sqrt{\sum_{i=1}^m c_i^2} = \|\mathbf{c}\| = 1$) such that

$$\sum_{i=1}^m c_i \mathbf{x}_i \simeq \boldsymbol{\varepsilon} \quad (9.7)$$

where $\boldsymbol{\varepsilon}$ is a “small” vector in the sense that $\|\boldsymbol{\varepsilon}\|$ is small.

The size of $\|\boldsymbol{\varepsilon}\|$ indicates the degree of the approximate linear dependence among the columns of \mathbf{X} . Of course \mathbf{c} is not unique, with the relative sizes of the components in different \mathbf{c} 's indicating, perhaps, qualitatively different dependencies. Large (relative to one) values of c_i indicate which variables are primarily involved in a particular approximate collinearity.

Since (9.7) is not a computationally convenient criterion to work with, many supposedly equivalent definitions of multicollinearity have been given [8]. However, as indicated in BKW, many of these conditions are neither necessary nor sufficient for (9.7) to hold, but not both. On the other hand, examination of the eigenvalues of $\mathbf{X}^T \mathbf{X}$ does.

To show this we consider the SVD of \mathbf{X} as described in Section 4.9. If $\mathbf{u}_i, 1 \leq i \leq n$, are the right singular vectors of \mathbf{X} , then

$$\mathbf{X} \mathbf{u}_i = \mu_i \mathbf{v}_i \quad (9.8)$$

where \mathbf{v}_i is the i -th left singular vector of \mathbf{X} and μ_i is the i -th right singular value of \mathbf{X} . Now if μ_i is small, then $\|\mu_i \mathbf{v}_i\| = \mu_i \|\mathbf{v}_i\| = \mu_i$, since $\|\mathbf{v}_i\| = 1$. Letting $\boldsymbol{\varepsilon} = \mu_i \mathbf{v}_i$ we find that

$$\mathbf{X} \mathbf{u}_i = \boldsymbol{\varepsilon}. \quad (9.9)$$

Letting $c_j = (\mathbf{u}_i)_j$, the j -th component of \mathbf{u}_i , (9.9) becomes

$$\sum_{j=1}^m c_j \mathbf{x}_j = \boldsymbol{\varepsilon} \quad (9.10)$$

where $\|\mathbf{c}\| = 1$ because \mathbf{u}_i is the i -th normalized eigenvector of $\mathbf{X}^T \mathbf{X}$. Since $\mu_i = \sqrt{\lambda_i}$, where λ_i is the i -th eigenvalue of $\mathbf{X}^T \mathbf{X}$, our basic multicollinearity diagnostic is to examine the eigenvalues of $\mathbf{X}^T \mathbf{X}$. Eigenvalues near zero indicate near dependencies in the data. In fact, (9.10) gives more, it tells us quantitatively what these dependencies are.

Conversely, if (9.10) holds then $\mathbf{X}^T\mathbf{X}$ has a small eigenvalue. For this we observe from linear algebra that if λ_{\min} is the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$, then

$$\begin{aligned}\lambda_{\min} &= \min_{\{\|\mathbf{d}\|=1\}} \langle \mathbf{d}, \mathbf{X}^T\mathbf{X}\mathbf{d} \rangle = \min_{\{\|\mathbf{d}\|=1\}} \langle \mathbf{X}\mathbf{d}, \mathbf{X}\mathbf{d} \rangle \\ &\leq \langle \mathbf{X}\mathbf{c}, \mathbf{X}\mathbf{c} \rangle = \|\mathbf{X}\mathbf{c}\|^2 = \|\boldsymbol{\varepsilon}\|^2\end{aligned}\quad (9.11)$$

where \mathbf{c} is given by (9.7). Thus,

$$\lambda_{\min} \leq \|\boldsymbol{\varepsilon}\|^2 \quad (9.12)$$

so that if $\|\boldsymbol{\varepsilon}\|$ is small, then $\mathbf{X}^T\mathbf{X}$ has at least one small eigenvalue, λ_{\min} .

This analysis indicates that a multicollinearity exists if and only if $\mathbf{X}^T\mathbf{X}$ has at least one small eigenvalue. These can be found by computing the spectral decomposition of $\mathbf{X}^T\mathbf{X}$. However, for numerical accuracy and stability it is generally preferable to compute the SVD of \mathbf{X} by an algorithm such as that given in [112].

Summarizing, multicollinearity exists if and only if $\mathbf{X}^T\mathbf{X}$ has at least one small eigenvalue λ . For each small eigenvalue a dependency is obtained by

$$\sum_{i=1}^n c_i \mathbf{x}_i \simeq 0 \quad (9.13)$$

where $\mathbf{c} = (c_1, c_2, \dots, c_m)^T$ is a normalized eigenvector of $\mathbf{X}^T\mathbf{X}$ corresponding to λ .

The question now arises as to how small an eigenvalue needs to be in order for a corresponding approximate linear dependency to exist. Again, this is at least a partly subjective matter, but some guidelines can be given from numerical analytical considerations in solving the normal equations $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$.

As indicated in Section 4.9 the condition number, $k(\mathbf{X}^T\mathbf{X}) = \mu_{\max}/\mu_{\min}$ governs the numerical stability of solving the normal equations. Since the singular values of $\mathbf{X}^T\mathbf{X}$ are the square roots of the eigenvalues of $(\mathbf{X}^T\mathbf{X})^T \mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X} = (\mathbf{X}^T\mathbf{X})^2$, they are the squares of the singular value of $\mathbf{X}^T\mathbf{X}$, then

$$\kappa(\mathbf{X}^T\mathbf{X}) = \lambda_{\max}/\lambda_{\min} \quad (9.14)$$

where λ_{\max} and λ_{\min} are the maximum and minimum eigenvalues of $\mathbf{X}^T\mathbf{X}$.

In BKW [8] they suggest that much real data may be known at best to four significant figures (since their interest was primarily in economics problems many of their observations and thus recommendations are drawn from an intimate knowledge of that type of data - they may not be universally valid [8]). If one wishes to preserve at least one significant figure in the solution, then one should tolerate condition numbers no larger than 1000. As a consequence, they advise that if the condition number exceeds 1000, then a multicollinearity problem may be present. This of course is equivalent to having $\lambda_{\max}/\lambda_{\min} > 1000$ or $\lambda_{\min} < \lambda_{\max}/1000$ as an indication as to how small λ_{\min} can be for $\mathbf{X}^T\mathbf{X}$ to be ill-conditioned.

If we define the *condition number* of \mathbf{X} as

$$\hat{\kappa}(\mathbf{X}) = \sqrt{\kappa(\mathbf{X}^T\mathbf{X})} \quad (9.15)$$

then a condition number of \mathbf{X} greater than 30 is considered to be *potentially damaging* to the regression analysis. As a rough guideline, all eigenvalues λ of $\mathbf{X}^T\mathbf{X}$ such that $\lambda_{\max}/\lambda > 1000$ indicate that a corresponding near dependency exists in the data.

Example 9.1 We consider the drink delivery data in Example in 5.14 to check for the possibility of multicollinearity problem in the data. First, we find the sample correlation matrix of (y, x_1, x_2) , \mathbf{R} is

$$\mathbf{R} = \begin{bmatrix} 1.0000 & 0.9646 & 0.8917 \\ 0.9646 & 1.0000 & 0.8242 \\ 0.8917 & 0.8242 & 1.0000 \end{bmatrix}$$

The standardized form of $\mathbf{X}^T\mathbf{X}$ is calculated as

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1.0000 & 0.7924 & 0.7891 \\ 0.7924 & 1.0000 & 0.9341 \\ 0.7891 & 0.9341 & 1.0000 \end{bmatrix}.$$

Then, setting $|\mathbf{X}^T\mathbf{X} - \lambda\mathbf{I}| = 0$ we obtain the eigenvalues

$$\lambda_1 = 2.6790, \lambda_2 = 0.2552, \lambda_3 = 0.06589.$$

Hence, the condition number is

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{2.6790}{0.0659} \simeq 40.65.$$

So we conclude that there is no serious multicollinearity problem in these data.

Example 9.2 Using the Hald cement data in Example in 5.16, we check for the potential of a multicollinearity problem in the data. First, we find the sample correlation matrix of (x_1, x_2, x_3, x_4, y) , \mathbf{R} is

$$\mathbf{R} = \begin{bmatrix} 1.0000 & 0.2286 & -0.8241 & -0.2455 & 0.7307 \\ 0.2286 & 1.0000 & -0.1392 & -0.9730 & 0.8163 \\ -0.8241 & -0.1392 & 1.0000 & 0.0295 & -0.5347 \\ -0.2455 & -0.9730 & 0.0295 & 1.0000 & -0.8213 \\ 0.7307 & 0.8163 & -0.5347 & -0.8213 & 1.0000 \end{bmatrix}$$

From this we observe that (x_1, x_3) , (x_1, x_4) , and (x_2, x_4) have a strong linear dependency. From the original $\mathbf{X}^T\mathbf{X}$, we have the normalized form of $\mathbf{X}^T\mathbf{X}$, which is

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1.0000 & 0.79715 & 0.95496 & 0.88617 & 0.88136 \\ & 1.0000 & 0.80216 & 0.47584 & 0.63255 \\ & & 1.0000 & 0.82713 & 0.70537 \\ & & & 1.0000 & 0.78750 \\ & & & & 1.0000 \end{bmatrix}.$$

The four eigenvalues for the Hald cement data are

$$\lambda_1 = 4.1196, \lambda_2 = 0.5539, \lambda_3 = 0.2887, \lambda_4 = 0.03768, \lambda_5 = 0.00009.$$

Hence, the condition number is

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{4.11960}{0.00009} \simeq 45,773.33,$$

which implies that there exists a serious multicollinearity problem in these data.

For further analysis of the data, we also present the matrix of corresponding eigenvectors \mathbf{e}_i .

$$\begin{aligned} \mathbf{e} &= [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5] \\ &= \begin{bmatrix} -0.49207 & -0.02282 & 0.03065 & 0.22485 & -0.84015 \\ -0.40060 & 0.76625 & -0.13253 & -0.47775 & 0.08112 \\ -0.46744 & 0.14241 & 0.51703 & 0.55101 & 0.43623 \\ -0.43481 & -0.56697 & 0.31671 & -0.61265 & 0.11765 \\ -0.43570 & -0.26572 & -0.78350 & 0.20557 & 0.28884 \end{bmatrix}. \end{aligned}$$

From our previous analysis the eigenvectors can be used to obtain the approximate collinearities that exist in the data. We first calculate the condition indices

$$\kappa_i = \lambda_{\max}/\lambda_i, \quad 1 \leq i \leq 5. \quad (9.16)$$

Those eigenvectors for which $\kappa_i > 1,000$ then indicate the approximate collinearities. In fact, if $\mathbf{e}_i = (e_{i1}, e_{i2}, \dots, e_{i5})$ is an eigenvector for which $\kappa_i > 1,000$, then an approximate collinearity is given by

$$\sum_{j=1}^n e_{ij}x_i \simeq 0. \quad (9.17)$$

Now

$$\begin{aligned} \kappa_1 &= 1, & \kappa_2 &= \frac{4.1196}{0.5539} = 7.437, \\ \kappa_3 &= \frac{4.1196}{0.2887} = 14.269, & \kappa_4 &= \frac{4.1196}{0.03768} = 109.33, \\ \kappa_5 &= \frac{4.1196}{0.00009} = 45,773.33. \end{aligned}$$

Hence, κ_5 is the only condition index which exceeds 1,000. Thus an approximation collinearity

$$-0.84015 + 0.08112x_1 + 0.43623x_2 + 0.11765x_3 + 0.28884x_4 \simeq 0 \quad (9.18)$$

exists in the data. Dropping the coefficient of x_1 we obtain the approximate collinearity

$$-0.84015 + 0.43623x_2 + 0.11765x_3 + 0.28884x_4 \simeq 0. \quad (9.19)$$

This suggests that x_1 should be included in the model, but at least one of (x_2, x_3, x_4) is superfluous. This is consistent with our variable selection results in Chapters 5 and 8.

9.3 Other Multicollinearity Diagnostics

Although it is now generally recognized that the eigenvalue structure of $\mathbf{X}^T\mathbf{X}$ is the most reliable way of detecting multicollinearity, historically many other methods have been proposed and a number are in common use today, even though they may be defective in one or more ways. In particular, because the eigenvalues of $\mathbf{X}^T\mathbf{X}$ do not have a simple statistical interpretation and may themselves be difficult to compute, these other measures can often be useful adjuncts, if not a replacement for the eigenvalue analysis of $\mathbf{X}^T\mathbf{X}$. A number of things that one frequently looks for are the following:

(i) Large correlations between the regressor variables

Typically, correlations exceeding 0.9 indicate the possible existence of multicollinearity problems. Intuitively, a correlation of this size between two columns of \mathbf{X} suggests that these columns are approximately linearly related. However, a high correlation between two variables may exist because these variables are implicated in a near dependency with other variables [8], so even though large correlations indicate the presence of multicollinearity, they do not necessarily indicate the full nature of the dependencies.

On the other hand, strong (even exact) near dependencies may exist in \mathbf{X} with all the pairwise correlations being small. Thus examining the correlation matrix $(\mathbf{X}_{sc}^* \mathbf{X}_{sc}^*)$ of the original regressors is not a totally satisfactory way of detecting multicollinearity.

As an example of this latter phenomenon, suppose we have $m-1$ independent random variables X_1, X_2, \dots, X_{m-1} such that $E(X_i) = 0, 1 \leq i \leq m-1$, and $Var(X_i) = 1, 1 \leq i \leq m-1$. Let $X_m = \sum_{j=1}^{m-1} X_j$, then $Cov(X_i, X_j) = 0, i \neq j, 1 \leq i, j \leq m-1$ and $Cov(X_i, X_m) = Var(X_i)$. Since $\sigma(X_m) = \sqrt{m-1}$, the correlation matrix ρ of X_1, X_2, \dots, X_m is given by

$$\rho = \begin{bmatrix} 1 & 0 & \cdots & 0 & \frac{1}{\sqrt{m-1}} \\ 0 & 1 & \cdots & 0 & \frac{1}{\sqrt{m-1}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \frac{1}{\sqrt{m-1}} \\ \frac{1}{\sqrt{m-1}} & \frac{1}{\sqrt{m-1}} & \cdots & \frac{1}{\sqrt{m-1}} & 1 \end{bmatrix}. \quad (9.20)$$

Obviously if m is large, all the correlations will be small, yet ρ is singular due to the linear relation $X_m = \sum_{j=1}^{m-1} X_j$.

(ii) Large values of the variance inflation factors

A typical guideline, as discussed in Chapter 5, is to regard any $VIF \geq 10$ to be possibly damaging to the analysis. This procedure is certainly sensible in view of the expression of the VIF in terms of the reciprocals of the eigenvalues of $\mathbf{X}^T \mathbf{X}$. In addition, the relation

$$VIF_i = \frac{1}{1 - R_i^2} \quad (9.21)$$

where R_i^2 is the multiple correlation of the regression of \mathbf{x}_i on $\mathbf{x}_j, j = 1, 2, \dots, m, i \neq j$, shows that a large value of any VIF_i corresponds to values of R_i^2 close to one. This, in turn indicates a possible approximate linear relation between the regressors. (note that a value $\delta_i \geq 10$ corresponds to $R_i^2 \geq 0.9$). However, since we have already noted that it is not always easy to relate the “strength” of a possible linear relationship to the size of R_i^2 , BKW find some fault with this indicator because of its lack of precision. On the other hand, VIFs have an easily understood statistical interpretation and frequently provide a reliable indicator of the presence of one or more collinearities. Since they are routinely computed when the regression is done in correlation form, they are readily available and should be examined in any regression analysis. In particular, if one is using a package where eigenvalue analysis is not available, they make a convenient, and usually reliable substitute.

(iii) Small values of $\det(\mathbf{X}^T \mathbf{X})$

This is a useful and again readily available diagnostic, but suffers from the deficiency of not being able to delineate the nature of the near dependencies. However, if this is not of interest then examining $\det(\mathbf{X}^T \mathbf{X})$ can be a reasonable alternative to a complete eigenvalue analysis.

(iv) Large estimated regression coefficients

The rationale for this diagnostic stems from the following observation. Consider $E(\langle \hat{\beta}, \hat{\beta} \rangle)$, the expected square length of $\hat{\beta}$. Then,

$$\begin{aligned} E(\langle \hat{\beta}, \hat{\beta} \rangle) &= \sum_{j=1}^m E(\hat{\beta}_j^2) \\ &= \sum_{j=1}^m \left\{ E(\hat{\beta}_j^2) - [E(\hat{\beta}_j)]^2 \right\} + \sum_{j=1}^m [E(\hat{\beta}_j)]^2. \end{aligned} \quad (9.22)$$

But, $E(\hat{\beta}_j) = \beta_j$ and $E(\hat{\beta}_j^2) - [E(\hat{\beta}_j)]^2 = \text{Var}(\hat{\beta}_j)$, so that

$$E(\langle \hat{\beta}, \hat{\beta} \rangle) = \sum_{j=1}^m \text{Var}(\hat{\beta}_j) + \langle \beta, \beta \rangle. \quad (9.23)$$

Now, as shown in (5.155)-(5.156) $\sum_{j=1}^m \text{Var}(\hat{\beta}_j) = \sigma^2 \text{tr}[(\mathbf{X}^T \mathbf{X})^{-1}] = \sigma^2 (\sum_{j=1}^m 1/\lambda_j)$, which gives

$$E(\langle \hat{\beta}, \hat{\beta} \rangle) = \langle \beta, \beta \rangle + \sum_{j=1}^m \frac{\sigma^2}{\lambda_j}. \quad (9.24)$$

Thus, if $\mathbf{X}^T \mathbf{X}$ has a small eigenvalue, $\langle \hat{\beta}, \hat{\beta} \rangle$ on average will be much larger than the true value $\langle \beta, \beta \rangle$. Thus the presence of multicollinearity tends to inflate the estimated length of $\langle \beta, \beta \rangle$ which suggests that at least one of $|\hat{\beta}_j|$, $1 \leq j \leq m$ will be large.

(v) Large standard errors of $\hat{\beta}$

As noted in Chapter 5,

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^n \frac{q_{ij}^2}{\lambda_j} \quad (9.25)$$

so that a small eigenvalue λ_j of $\mathbf{X}^T \mathbf{X}$ may lead to a large value of this quantity. This generally indicates that the estimation of β_j is unstable and that considerably different values would be obtained if the data were slightly altered. Statistically, this leads to wide confidence intervals for β_j and a corresponding insignificant observed t value. Since the presence of just one approximate dependency may inflate the variances of all of the coefficient estimates, one may find that many or all of the estimated coefficients fail to appear significantly different from zero even if the overall fit is excellent. This phenomenon was seen in the Hald data in Example 5.16 and the Longley data in Example 5.17.

(vi) One or more estimated regression coefficients of the wrong sign

Often in practice a theory and/or common sense suggests what the sign of a regression coefficient ought to be. For example, if one was building a salary model for a particular population one would anticipate that “years of service” would be a significant variable with a positive coefficient. Estimation of this coefficient from a particular set of data leading to a negative coefficient would surely cause one to become concerned. However, if the model also had “age” as a variable then multicollinearity due to these variables might be anticipated leading to possible large standard errors of the coefficient for “years of service”. The resultant instability in estimation could then lead to the wrong sign for this coefficient.

A detailed discussion of this phenomenon may be found in BKW [8]. On the other hand, a wrong sign may be due to bias resulting from misspecification of the model or to large values of σ^2 , rather than the problem data.

In summary, it is wise to keep in mind that multicollinearity is a problem with the design matrix and one should diagnose its presence using measures which depend only on \mathbf{X} and not on quantities which involve the dependent variable Y . In this regard, in examining the sizes, signs and t values of $\hat{\beta}_j$, $1 \leq j \leq m$, one is really looking more at the consequences of multicollinearity rather than determining its presence. In this regard, the eigenvalues of $\mathbf{X}^T\mathbf{X}$, the variance inflation factors and $\det(\mathbf{X}^T\mathbf{X})$ are useful for this purpose, while $\hat{\beta}$ and various statistical calculated quantities may not be.

9.3.1 Consequences of Multicollinearity

We have already discussed the consequences of multicollinearity on the least squares estimation of β in a number of places. So here we merely summarize our previous observations. First, multicollinearity makes $\mathbf{X}^T\mathbf{X}$ highly ill-conditioned and this can lead to large round-off errors in the numerical calculation of $\hat{\beta}$ for any method that uses the normal equations to compute $\hat{\beta}$. To a large extent, this problem can be alleviated by using techniques such as the QR decomposition of \mathbf{X} which avoids forming $\mathbf{X}^T\mathbf{X}$. In addition, with today’s computers which have high precision, this problem is less serious than in the past.

Second, estimation of at least some components of β may be unstable due to the presence of multicollinearity. (This instability may often be demonstrated by observing substantial changes in the coefficient estimates if a variable or observation is deleted.)

This may give estimated coefficients with the wrong sign and/or too large a magnitude. In addition, many variables may have small t values leading to the apparent contradiction of having an overall good fit with very few or no significant coefficients. In this regard, it is of some interest to be able to determine a priori which coefficients may be estimated poorly and we digress briefly to discuss a further diagnostic procedure introduced by BKW for this purpose.

Let

$$\pi_{ij} = \frac{q_{ij}^2/\lambda_i}{VIF_j}, \quad j = 1, 2, \dots, (m+1) \quad (9.26)$$

denote the proportion of the variance of $\hat{\beta}_j$ due to λ_i . (This is called the *variance decomposition proportion* by BKW.) If for fixed i , π_{ij} is large for two or more values of j , then λ_i is contributing a large proportion of the variance of those variables. If λ_i is an eigenvalue responsible for an approximate dependency, then these variables are

implicated in the dependency and that dependency will generally degrade the estimation of the corresponding β_j 's. As a guideline BKW suggest that if $\pi_{ij} > 0.5$ then π_{ij} should be considered large for any λ_i whose *condition index* $\mu_i/\mu_{\max} > 30$. (This diagnostic should produce essentially the same information as an examination of the sizes of the coefficients of the eigenvector(s) corresponding to λ_i .)

Further difficulties arise in variable selection as has been pointed out in Chapter 8.

9.3.2 Prediction

In contrast, to estimation, multicollinearity need not cause problems with using the model for prediction, so long as one predicts at points which are consistent with the collinearities in the original data. To elaborate, we suppose that the model fits the data overall and we want to use the model to estimate the response Y as some point \mathbf{x}_0 . As usual, the precision of this estimate will be measured by $\text{Var}(\hat{Y}_{\mathbf{x}_0})$ which according to Eq. (5.286) is estimated by

$$\widehat{\text{Var}}(Y_{\mathbf{x}_0}) = s^2 \langle \mathbf{x}_0, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T \rangle. \quad (9.27)$$

We analyze this further using the spectral decomposition of $(\mathbf{X}^T \mathbf{X})^{-1}$. From (4.125) $(\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^T$ so that

$$\begin{aligned} \langle \mathbf{x}_0, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0^T \rangle &= \langle \mathbf{x}_0, \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^T \mathbf{x}_0^T \rangle = \langle \mathbf{Q}^T \mathbf{x}_0, \mathbf{\Lambda}^{-1} \mathbf{Q}^T \mathbf{x}_0^T \rangle \\ &= \langle \mathbf{v}, \mathbf{\Lambda}^{-1} \mathbf{v} \rangle = \sum_{j=1}^m \frac{v_j^2}{\lambda_j}, \end{aligned} \quad (9.28)$$

where $\mathbf{v} = (v_1, v_2, \dots, v_m)^T = \mathbf{Q}^T \mathbf{x}_0$.

Now suppose for sake of argument that λ_i is a small eigenvalue which contributes a near dependency \mathbf{X} . The corresponding component v_i^2/λ_i in (9.28) will be large, unless the effect of $1/\lambda_i$ is counterbalanced by a small value of v_i . For this to happen, we observe that

$$v_i = (\mathbf{Q}^T \mathbf{x}_0)_i = \sum_{j=1}^m q_{ji} x_{0j} \quad (9.29)$$

where $\mathbf{Q} = [q_{ij}]$.

Now (q_{ji}) , $1 \leq j \leq m$, is the i -th column of \mathbf{Q} which is the i -th eigenvector of $\mathbf{X}^T \mathbf{X}$. Recalling that this eigenvector defines a collinearity, it follows that if \mathbf{x}_0 satisfies this collinearity, then $v_i \simeq 0$. Thus, if \mathbf{x}_0 is consistent with the collinearity, determined by λ_i then v_i^2/λ_i may be small even if $1/\lambda_i$ is large. Hence, if \mathbf{x}_0 is consistent with all the collinearities in the original data, then $\widehat{\text{Var}}(Y_{\mathbf{x}_0})$ may still be small in the presence of small eigenvalues.

One can think of this phenomenon in geometric terms as follows. Roughly speaking, one can think that multicollinearities in the data restrict the columns of \mathbf{X} to lie in a subspace of the full m -dimensional column space of \mathbf{X} . Thus prediction can be reliable for points in this subspace but generally unreliable for points orthogonal to this space.

With the above argument in mind, some authors take the view that if severe multicollinearity is present one should not concern oneself with estimation, because the data

may be non-informative in this respect. The model may be used for prediction, but only at points consistent with the collinearities in the original data. Since the observed data are unable to provide one with information about the nature of the model outside of this region. Apparently this is a view held by only a minority, if the recent explosion of research on “improved” estimation methods seems to indicate [26]. We now turn our attention to this subject.

9.4 Combatting Multicollinearity

Suppose that one has concluded that one or more strong multicollinearities are present in the data and one is interested in estimation and prediction at points not necessarily in line with the observed near dependencies. What, if anything, can one do? Ignoring computational problems, the primary effect of multicollinearity will be to produce large estimation and/or prediction errors if least squares estimation is used. Since the Gauss-Markov theorem says that the least squares estimator is the best linear unbiased estimator (BLUE) of β , reduction in variance will then be feasible only if we use some *nonlinear estimator* or a *biased linear estimator*. Both of these approaches have been utilized, with current statistical thought seeming to favor biased linear estimation. It is this approach that we shall discuss in some detail. However, before resorting to such an approach, it is worthwhile to examine what remedies may be available in the context of classical estimation.

First, coefficients of variables not involved in any near dependency may not be affected by multicollinearity. If one is only interested in these coefficients then no remedy need be taken. However, in most cases estimation of all of the coefficients will be of concern so that this circumstance is probably mostly of theoretical interest.

More generally, certain linear combinations of the regression coefficients may be estimated accurately even if the individual coefficients cannot. Using the model for prediction is a particular case of this and as has been pointed out in the previous Section, certain predictions may require no remediation.

As Belsley, Kuh and Welsch (BKW) [8] emphasize, multicollinearity is a problem with the design matrix \mathbf{X} , and is not a statistical problem, so initial strategies might focus on modifying \mathbf{X} if at all possible. In this regard the following approaches have been suggested.

(i) Collect new data

If multicollinearity is present because of the way the data were collected, then collecting new data may solve the problem. For instance, collecting data in a narrow range of any regressor variable \mathbf{x} may introduce an approximate linear dependence with the intercept. As a particular example, in economics, regression models are often used to model “total consumption” in a given year as a function of previous years consumption, “total income” and other possible variables such as interest rates [8]. Until recently, interest rates varied over rather narrow ranges and when used in a regression model were often found to be statistically insignificant as indicated by the usual t -test. Because of the approximate linear dependence of the interest rate variable and the intercept, the problem of whether to include an interest rate variable in these equations has been a source of much controversy [8]. In this regard, the large variations in interest rates during the 70’s, 80’s

and 90's may provide useful new data to economists even if they were not beneficial to consumers.

In many cases new data cannot be collected and the analyst is stuck with what he has. In addition, if the variables are approximately collinear, by definition, it may not be possible to modify \mathbf{X} in any simple way.

(ii) Model respecification

Changing the variables may help. For instance in polynomial regression, artificial collinearities may be introduced by using uncentered variables - centering the variables may readily solve the problem.

Eliminating one or more of the variables present in any near dependency is a classical solution, and variable selection techniques as discussed in Chapter 8 are often employed in this regard. However, such procedures may seriously bias the resultant model and result in poor predictive models, particularly if the multicollinearity was caused by the choice of variables in available data, and are not due to any time dependency of the underlying variables. (Note that in reality such techniques are really biased estimation procedures but often not discussed in that context.)

In this regard the technique of using “*auxiliary regressions*” following the delineation of the multicollinearities using the techniques of BKW [8] may be of value as an alternative to standard variable selection techniques.

If none of these approaches is feasible, then attention usually shifts to improving the estimation using the data at hand.

9.5 Biased Estimation

9.5.1 Shrunken Estimators

Since we have seen that one of the degrading effects of multicollinearity is to produce “large” unstable estimates of β , it is reasonable to focus on biased estimation procedures which try to shrink the size of the least squares estimator of β_j and/or its variance. As we pointed out previously, due to the Gauss-Markov theorem, biased estimation is necessary in this regard if we wish to stay in the framework of linear estimation.

The idea is similar in philosophy to that used in variable selection (which we have already indicated is a possible solution to the multicollinearity problem), that is, hopefully, we can trade a small bias in estimation for a large reduction in variance. This suggests that we consider the class of estimators (B for biased)

$$\hat{\beta}_B = \mathbf{A}\hat{\beta} \quad (9.30)$$

where \mathbf{A} is an $m \times m$ unknown matrix and $\hat{\beta}$ is the least squares estimator of β . \mathbf{A} then has to be chosen to produce a desirable trade-off between bias and variance.

As we did in the variable selection problem, this is usually done by picking \mathbf{A} to minimize the MSE of $\hat{\beta}_B$. Since choosing $\mathbf{A} = \mathbf{I}_m$ gives the least squares estimator, the minimizing value of \mathbf{A} (if it exists) will have a MSE no larger than $\text{Var}(\hat{\beta})$. Thus, this approach appears a reasonable way to solve the problem. On the other hand, we have no guarantee that such constants exist. We begin by assuming that A is a scalar so that $\hat{\beta}_B = c\hat{\beta}$.

In this case

$$E(\hat{\beta}_B) = cE(\hat{\beta}) = c\beta \quad (9.31)$$

and

$$Var(\hat{\beta}_{B_i}) = c^2 Var(\hat{\beta}_i) = c^2 \sigma^2 \delta_i. \quad (9.32)$$

Thus the bias vector is $\beta - c\beta = (1 - c)\beta$ and

$$MSE(\hat{\beta}_B) = Var(\hat{\beta}_B) + (Bias)^2 \quad (9.33a)$$

$$= c^2 \sigma^2 \sum_{i=1}^m \delta_i + (1 - c)^2 \sum_{i=1}^m \beta_i^2. \quad (9.33b)$$

Minimizing $MSE(\hat{\beta}_B)$ with respect to c gives the optimum value of c as

$$c_{opt} = \frac{\sum_{i=1}^m \beta_i^2}{\sum_{i=1}^m \beta_i^2 + \sigma^2 \sum_{i=1}^m \delta_i}. \quad (9.34)$$

From this we see that $0 \leq c_{opt} \leq 1$ so that $\hat{\beta}$ is shrunk towards zero and an easy calculation shows that $MSE(\hat{\beta}_B) \leq Var(\hat{\beta})$. Hence, c_{opt} gives an estimator which appears to counteract some of the deleterious effects of multicollinearity. However, in examining $c_{opt}\hat{\beta}$ some problems are evident. First, using $c_{opt}\hat{\beta}$ shrinks all coefficients equally and a priori there is no reason why this should be done. Similarly, if we want an estimator that can correct incorrect signs then $c_{opt}\hat{\beta}$ is unable to do this.

A further difficulty (and this is characteristic of virtually all currently used biased estimators) is that c_{opt} depends on β and σ^2 - which are unknown - so that strictly speaking $c_{opt}\hat{\beta}$ is not an estimator. If we now estimate c_{opt} by

$$\hat{c}_{opt} = \frac{\sum_{i=1}^m \hat{\beta}_i^2}{\sum_{i=1}^m \hat{\beta}_i^2 + s^2 \sum_{i=1}^m \delta_i} \quad (9.35)$$

and use $\hat{c}_{opt}\hat{\beta}$ to estimate β , then $\hat{c}_{opt}\hat{\beta}$ is an estimator but now we have no guarantee that this estimator has smaller MSE than $\hat{\beta}$. (In case $\mathbf{X}^T \mathbf{X} = \mathbf{I}_m$, there are estimators of the form $\hat{c}\hat{\beta}$, where \hat{c} depends on $\hat{\beta}$ and s^2 , which are known to have smaller MSE than $\hat{\beta}$. This is the famous *James-Stein estimator* [67]. Further discussion can be found in [32, 33, 118].)

To remedy the problem of equal shrinkage we could use m different shrinking factors $c_i, 1 \leq i \leq m$, and the estimator $(c_1\hat{\beta}_0, c_2\hat{\beta}_0, \dots, c_m\hat{\beta}_0)^T$. The c_i 's can be chosen to minimize the mean square error given by

$$\sigma^2 \sum_{i=1}^m c_i^2 \delta_i + \sum_{i=1}^m (1 - c_i)^2 \beta_i^2, \quad 1 \leq i \leq m. \quad (9.36)$$

Doing this gives

$$c_i^{opt} = \frac{\beta_i^2}{\beta_i^2 + \sigma^2 \delta_i}, \quad 1 \leq i \leq m. \quad (9.37)$$

If we replace c_i^{opt} with

$$c_i^{opt} = \frac{\hat{\beta}_i^2}{\hat{\beta}_i^2 + s^2 \delta_i}, \quad (9.38)$$

again we arrive at a true shrunken estimator of β which again is not guaranteed to minimize the MSE.

Since these estimators seem not to be used in practice, we will discuss them no further. Rather, the reader can view this section as a prelude to the basic ideas of ridge estimators which shrink $\hat{\beta}$ in a much more complicated way.

9.5.2 Ridge Regression

Although shrunken estimators appear reasonable as a way of combating at least some consequences of multicollinearity, they seem not to be used in practice, perhaps because they take no direct account of what happens to the fit in terms of attempting to minimize *SSE*. Thus, another point of view towards shrinking $\hat{\beta}$ is to choose $\hat{\beta}_B$ so that it minimizes *SSE* subject to the constraint that $\langle \hat{\beta}_B, \hat{\beta}_B \rangle < \langle \hat{\beta}, \hat{\beta} \rangle$. In this case, we try to find $\hat{\beta}_B$ as that which solves the constrained minimization problem:

$$\begin{aligned} & \underset{\beta}{\text{Minimize}} \quad SSE = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle \\ & \text{subject to : } \langle \hat{\beta}_R, \hat{\beta}_R \rangle = d^2 \leq \langle \hat{\beta}, \hat{\beta} \rangle, \quad (R \text{ for ridge}) \end{aligned} \quad (9.39)$$

where $d \geq 0$ is a prior restriction on the length of the parameter vector β . On the other hand, it can be shown that shrunken estimators of the form $c\hat{\beta}$ are solutions to the minimization problem

$$\begin{aligned} & \underset{\beta}{\text{Minimize}} \quad \langle \mathbf{V}(\mathbf{y} - \mathbf{X}\beta), \mathbf{V}(\mathbf{y} - \mathbf{X}\beta) \rangle \\ & \text{subject to : } \langle \beta, \beta \rangle = d^2 \end{aligned} \quad (9.40)$$

where $\mathbf{V} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}$. Since the criterion function $\langle \mathbf{V}(\mathbf{y} - \mathbf{X}\beta), \mathbf{V}(\mathbf{y} - \mathbf{X}\beta) \rangle$ is not the *SSE*, the solution to (9.39) seems to be more appropriate.

As we show next, the solution to (9.40) leads to the so called *ridge estimators* which have received much attention as possible “cures” for the multicollinearity problem [?]. We begin by showing that the solution to the minimization problem (9.40) satisfies the ridge estimation equation

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m) \hat{\beta}_R = \mathbf{X}^T \mathbf{y} \quad (9.41)$$

where λ is chosen to satisfy the constraint

$$\langle \hat{\beta}_R, \hat{\beta}_R \rangle = d^2. \quad (9.42)$$

To prove this, we again use the technique of Lagrange multipliers [104].

Thus, let

$$L = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle + 2\lambda (\langle \beta, \beta \rangle - d^2) \quad (9.43)$$

and then the minimizing values of (β, λ) are obtained by solving

$$\partial L / \partial \beta = 0, \quad \partial L / \partial \lambda = 0. \quad (9.44)$$

Doing these differentiations

$$\frac{\partial L}{\partial \beta} = 2\mathbf{X}^T \mathbf{X} \beta - 2\mathbf{X}^T \mathbf{y} + 2\lambda \beta = 0, \quad (9.45)$$

$$\frac{\partial L}{\partial \lambda} = \langle \beta, \beta \rangle - d^2 = 0, \quad (9.46)$$

and calling a solution to (9.43) $\hat{\beta}_R$, we see that (9.44)-(9.46) yields (9.41) and (9.42). We will now show that if $\lambda \geq 0$ and \mathbf{X} has full rank then (9.44)-(9.46) have a unique solution. This solution is called the *ridge estimate* of $\hat{\beta}$ and λ is called the *ridge parameter*. (Of course λ is a function of d .)

First, observe that for any $\lambda \geq 0$, $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m$ has an inverse (this was shown in Chapter 5 for $\lambda = 0$, the proof for $\lambda > 0$ is similar since it is easily shown that $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m$ is positive definite. Thus, (9.43) has a unique solution

$$\hat{\beta}_R = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{y}. \quad (9.47)$$

We now need to show that λ can be chosen to satisfy (9.47) for any $d \geq 0$. For this we will need an alternative expression for $\langle \hat{\beta}_R, \hat{\beta}_R \rangle$ which will be given in Theorem 9.1. That is,

$$\langle \hat{\beta}_R, \hat{\beta}_R \rangle = \frac{\sum_{i=1}^m \lambda_i \hat{\gamma}_i^2}{(\lambda + \lambda_i)^2} \quad (9.48)$$

where λ_i , $1 \leq i \leq m$, are the eigenvalues of $\mathbf{X}^T \mathbf{X}$ and $\hat{\gamma}_i = (\mathbf{Q}^T \hat{\beta})_i$, $1 \leq i \leq m$, where \mathbf{Q} is the matrix of normalized eigenvectors of $\mathbf{X}^T \mathbf{X}$.

Now if $\lambda = 0$, $\langle \hat{\beta}_R, \hat{\beta}_R \rangle = \sum_{i=1}^m \hat{\gamma}_i^2 / \lambda_i = \langle \hat{\beta}, \hat{\beta} \rangle$, while (9.48) shows that $\langle \hat{\beta}_R, \hat{\beta}_R \rangle$ is a decreasing function of λ for $\lambda > 0$. Additionally, as $\lambda \rightarrow \infty$, $\langle \hat{\beta}_R, \hat{\beta}_R \rangle \rightarrow 0$ so that if $d^2 \leq \langle \hat{\beta}, \hat{\beta} \rangle$ there is a unique value of $\lambda \geq 0$ satisfying $\langle \hat{\beta}_R, \hat{\beta}_R \rangle = d^2$.

If d^2 were known in advance then we could determine λ by solving the nonlinear equation

$$\frac{\sum_{i=1}^m \lambda_i \hat{\gamma}_i^2}{(\lambda + \lambda_i)^2} = d^2 \quad (9.49)$$

and the ridge estimator would then be given by

$$\hat{\beta}_R = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{y}. \quad (9.50)$$

Since, by definition $d^2 \leq \langle \hat{\beta}, \hat{\beta} \rangle$, the ridge estimator shrinks $\hat{\beta}$ towards the origin as do the estimators in the previous section. In this case, the shrinkage is accomplished (see Theorem 9.1) by multiplying the least squares estimator by the matrix

$$\mathbf{A}(\lambda) = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X}, \quad (9.51)$$

so it is an estimator of the form in (9.50).

Of course in practice d^2 will not be known and it will generally have to be chosen from the data to provide an estimator which one hopes is somehow "better" than $\hat{\beta}$.

However, as is emphasized by Draper and Van Nostrand [28] and Draper and Smith [27]. No matter how λ is chosen, the ridge estimate can always be viewed as a least squares estimator of β where prior information is used to constrain the length of the resulting estimator. (This interpretation of $\hat{\beta}_R$ has been used as a basis for criticism of many of the published simulations studies purposing to show that ridge regression is better than OLS [28].)

To get some intuitive feel for what ridge estimation does, recall that multicollinearity is generally indicated by a large condition number of $\mathbf{X}\mathbf{X}^T$. Since the eigenvalues of $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_m$ are $\lambda_i + \lambda$, $1 \leq i \leq m$, where λ_i , $1 \leq i \leq m$, are the eigenvalues of $\mathbf{X}\mathbf{X}^T$, choosing small values of λ can substantially improve the conditioning of $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_m$ over that of $\mathbf{X}^T\mathbf{X}$.

Example 9.3 Suppose that $\lambda_{\max} = 1$ and $\lambda_{\min} = 10^{-4}$. Then $\kappa(\mathbf{X}^T\mathbf{X}) = 10^4$. Choosing $\lambda = 10^{-2}$

$$\kappa(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_m) = \frac{1 + 10^{-2}}{10^{-4} + 10^{-2}} = \frac{1.01}{0.0101} = 100,$$

a reduction by a factor of 100. Similar reductions can be found for the VIFs.

It is this property of reduction in conditioning which was a major factor in introducing the technique of adding a small constant to the diagonal elements of an ill-conditioned matrix. It is interesting to note that this idea is widely used in many other areas of applied mathematics [41] where it is usually called *regularization*.

How to choose λ in the absence of realistic prior information on the length of $\hat{\beta}_R$ is the most difficult part of ridge estimation and has led to a large number of techniques [27, 28, 87] with different methods often providing conflicting results [39]. The difficulty, as with the shrunken estimators $c\hat{\beta}$, is that if $\hat{\beta}_R$ is chosen to have certain optimality properties, such as minimizing the MSE, then the ridge parameter λ will then depend on the unknown values of β . When these values are replaced by estimates, λ becomes a random variable $\hat{\lambda}$, with a generally unknown distribution, whose use in (9.50) gives rise to an estimator which is no longer known to have any optimality properties. Researchers have attempted to resolve some of these problems via Monte Carlo simulations.

However, as is shown in Gibbons (1981) [39] the resulting properties of the different ridge estimators is again heavily dependent on \mathbf{X} and β and σ^2 and do not always improve on OLS. This property (at least in the author's opinion) only serves to accentuate the criticism of Draper and Van Nostrand that ridge estimation always implies prior information concerning β . If such information is available and reliable, then the statistician may proceed in the spirit of Bayesian analysis to use this information in a rational way to (perhaps) improve on OLS. In the absence of such information, the use of ridge estimation (or any similar biased estimation method) will not be guaranteed to improve on OLS, and may in fact be worse.

Notwithstanding these caveats, we now turn to a presentation of some of the more popular methods for selecting λ . In this regard we have been guided by the simulation results of Gibbons [39] and the survey of Draper and Van Nostrand [28].

To motivate many of these choices we will need to know a number of properties of the ridge estimator $\hat{\beta}_R$ for a fixed value of λ . These are summarized in Theorem 9.1.

Theorem 9.1 (Properties of $\hat{\beta}_R$ when λ is known) Suppose that $\hat{\beta}_R$ is the unique solution of (9.39) with $\lambda \geq 0$, known. Then,

$$(i) \hat{\beta}_R = A(\lambda) \hat{\beta},$$

where $A(\lambda) = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X}$ and $\hat{\beta}$ is the least squares estimator of β .

(ii) $\hat{\beta}_R$ is generally a biased estimator of β if $\lambda > 0$ with the bias given by

$$E(\hat{\beta}_R) - \beta = -\lambda (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \beta. \quad (9.52)$$

(iii) The variance-covariance matrix $\Sigma(\hat{\beta}_R)$ of $\hat{\beta}_R$ is given by

$$\Sigma(\hat{\beta}_R) = \sigma^2 (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1}. \quad (9.53)$$

(iv) The mean square error of $\hat{\beta}_R$ is given by

$$MSE(\hat{\beta}_R) = \sum_{j=1}^m \sigma^2 \lambda_j / (\lambda_j + \lambda)^2 + \lambda^2 \sum_{j=1}^m \gamma_j^2 / (\lambda_j + \lambda)^2 \quad (9.54)$$

where $\lambda_j, 1 \leq j \leq m$, are the eigenvalues of $\mathbf{X}^T \mathbf{X}$, $\gamma_i = (\mathbf{Q}^T \beta)_i, 1 \leq i \leq m$, and \mathbf{Q} is the matrix of orthonormal eigenvectors of $\mathbf{X}^T \mathbf{X}$.

(v) (Hoerl and Kennard [58]). There exists a value of $\lambda > 0$ such that

$$MSE(\hat{\beta}_R) < MSE(\hat{\beta}). \quad (9.55)$$

(vi) If $\lambda > 0$,

$$\langle \hat{\beta}_R, \hat{\beta}_R \rangle < \langle \hat{\beta}, \hat{\beta} \rangle \quad (9.56)$$

Proof. (i) From (9.41)

$$\begin{aligned} \hat{\beta}_R &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{A}(\lambda) \hat{\beta}. \end{aligned} \quad (9.57)$$

(ii) Since $\hat{\beta}$ is unbiased,

$$E(\hat{\beta}_R) = \mathbf{A}(\lambda) E(\hat{\beta}) = \mathbf{A}(\lambda) \beta \neq \beta \quad (9.58)$$

unless $\mathbf{A}(\lambda) = \mathbf{I}_m$ for all $\beta \in \mathbb{R}^m$. And this requires $\lambda = 0$. The bias of $\hat{\beta}_R$ is then given by

$$\begin{aligned} E(\hat{\beta}_R) - \beta &= \mathbf{A}(\lambda) \beta - \beta = [(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} - \mathbf{I}_m] \beta \\ &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} [-\lambda \mathbf{I}_m] \beta = -\lambda (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \beta. \end{aligned} \quad (9.59)$$

(iii) Since $\hat{\beta}_R = \mathbf{A}(\lambda) \hat{\beta}$,

$$\begin{aligned} \Sigma(\hat{\beta}_R) &= \mathbf{A}(\lambda) \Sigma(\hat{\beta}) \mathbf{A}^T(\lambda) = \sigma^2 \mathbf{A}(\lambda) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{A}^T(\lambda) \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1}. \end{aligned} \quad (9.60)$$

(iv) From Equation (9.33a) we have

$$MSE(\hat{\beta}_R) = Var(\hat{\beta}_R) + (Bias)^2 \quad (9.61)$$

and using (9.60) and (9.61)

$$(Bias)^2 = \lambda^2 \left\langle (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \boldsymbol{\beta}, (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \boldsymbol{\beta} \right\rangle \quad (9.62)$$

and

$$\begin{aligned} Var(\hat{\beta}_R) &= \text{tr} \left[\Sigma(\hat{\beta}_R) \right] \\ &= \sigma^2 \text{tr} \left[(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \right] \\ &= \sigma^2 \text{tr} \left[(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-2} \mathbf{X}^T \mathbf{X} \right]. \end{aligned} \quad (9.63)$$

Now, $\mathbf{X}^T \mathbf{X} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T$ so that $(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T + \lambda \mathbf{I}_m = \mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m) \mathbf{Q}^T$ which gives

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-2} = \mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-2} \mathbf{Q}^T \quad (9.64)$$

and

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-2} \mathbf{X}^T \mathbf{X} = \mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-2} \boldsymbol{\Lambda} \mathbf{Q}^T. \quad (9.65)$$

Thus,

$$\begin{aligned} \text{tr} \left[(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-2} \mathbf{X}^T \mathbf{X} \right] &= \text{tr} \left[\mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-2} \boldsymbol{\Lambda} \mathbf{Q}^T \right] \\ &= \text{tr} \left[\mathbf{Q}^T \mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-2} \boldsymbol{\Lambda} \right] \\ &= \text{tr} \left[(\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-2} \boldsymbol{\Lambda} \right] \quad (\text{since } \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_m) \\ &= \sum_{i=1}^m \frac{\lambda_i}{(\lambda_i + \lambda)^2}. \end{aligned} \quad (9.66)$$

Using this in (9.63) gives the first term in (9.61).

Again, using the spectral decomposition of $\mathbf{X}^T \mathbf{X}$ and the orthogonality of \mathbf{Q}

$$\begin{aligned} (Bias)^2 &= \lambda^2 \left\langle \mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{Q}^T \boldsymbol{\beta}, \mathbf{Q} (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{Q}^T \boldsymbol{\beta} \right\rangle \\ &= \lambda^2 \left\langle (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{Q}^T \boldsymbol{\beta}, (\boldsymbol{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{Q}^T \boldsymbol{\beta} \right\rangle \\ &= \sum_{i=1}^m \frac{\lambda^2 \gamma_i^2}{(\lambda_i + \lambda)^2} \end{aligned} \quad (9.67)$$

where $\gamma_i = (\mathbf{Q}^T \boldsymbol{\beta})_i$. Finally, using (9.66) and (9.67) in the expression (9.61) for $MSE(\hat{\beta}_R)$ we get

$$MSE(\hat{\beta}_R) = \sum_{i=1}^m \frac{\lambda^2 \gamma_i^2 + \lambda_i \sigma^2}{(\lambda_i + \lambda)^2}. \quad (9.68)$$

(v) From (9.68) we see that if $\lambda = 0$, then $MSE(\hat{\beta}_R) = MSE(\hat{\beta})$ so that in order to prove (v) it suffices to prove that $MSE(\hat{\beta}_R)$ is a decreasing function near $\lambda = 0$ and for this all we need to show is that $\partial MSE(\hat{\beta}_R) / \partial \lambda|_{\lambda=0} < 0$. Taking this derivative we find that

$$\left. \frac{\partial MSE(\hat{\beta}_R)}{\partial \lambda} \right|_{\lambda=0} = -2\sigma^2 \sum_{i=1}^m \frac{\gamma_i^2}{\lambda_i^2} < 0. \quad (9.69)$$

(vi) The calculations for (9.54) are similar to those in proving (v) and are left as an exercise. ■

9.5.3 Choosing the Ridge Parameter

Each of the various properties of the ridge estimator that we have discussed so far has motivated one or more methods for estimating λ . Generally, these attempt to do one or more of the following:

- (i) Reduce the variance and/or stabilize the coefficient estimates relative to OLS.
- (ii) Produce an estimator with a smaller MSE or prediction mean squared error than OLS.

As part (v) of Theorem 9.1 shows, there exists a value of the ridge parameter λ which gives an estimator $\hat{\beta}_R$ with smaller MSE than for $\hat{\beta}$. Ideally, the best estimator from this point of view would be the one that minimizes the MSE. However, as (9.69) shows this minimum value would generally depend on β and σ^2 which are unknown and so such an estimator cannot be implemented in practice. One way out would be to work iteratively by first estimating β and σ^2 in (9.54), say by OLS, finding the value of λ that minimizes the resulting estimate of $MSE(\hat{\beta}_R)$ and repeat until these estimates appear to have converged. However, even if this were done, we would still have no guarantee (and we know of no proof) that the resulting estimator would have smaller MSE than the OLS estimate $\hat{\beta}$.

Typically, then, each procedure chosen yields a different value of λ , with these values differing substantially for any given problem depending on the method used. As a consequence, at the current stage of development of the subject it is not possible to say which procedure is “best” or even if any generally provide estimators which are better than OLS. However, since such methods are in widespread use, we will discuss a number of those which seem to have shown to have the most promise. Here we have to keep in mind that the relative merit of the various techniques have for the most part been studied by simulation, since very little is known theoretically. At the same time remember that some authors have found methodological faults with these studies as well.

The methods for choosing λ fall into two broad categories: *stochastic* and *nonstochastic*. Stochastic methods use the observed values \mathbf{y} in some way so that the resulting λ 's are actually random variables. Nonstochastic methods make use only of the design matrix \mathbf{X} . We begin with a discussion of stochastic methods, since these have proved to be the most popular in practice.

Stochastic Methods for Choosing λ

(1) The Ridge Trace This method, introduced by Hoerl and Kennard [58, 59] is perhaps the oldest and apparently the most recommended method for choosing λ although virtually no formal evidence can be found in its favor. Here the ridge estimator $\hat{\beta}_R(\lambda)$ is calculated for a sequence of values of λ , typically, $0 \leq \lambda \leq 1$, and the individual coefficients of $\hat{\beta}_R(\lambda)$ are plotted against λ . As observed previously, $\langle \hat{\beta}_R(\lambda), \hat{\beta}(\lambda) \rangle < \langle \beta, \beta \rangle$, $\lambda > 0$, and $\beta_R(\lambda) \rightarrow 0$, $\lambda \rightarrow \infty$, so we expect that these plots will show a rapid decline in the magnitude of these coefficients as λ increases and may tend to stabilize at some values of λ which is usually determined by visual inspection. Typical behavior is shown in Figure 9.1.

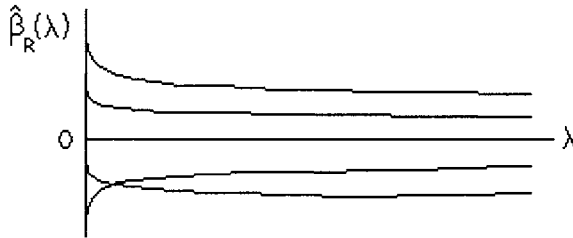


Figure 9.1: Graph of ridge trace

This method is highly subjective, since it is not exactly clear what value of λ represents stability and the fact that different plots stabilize at different points. (Perhaps plotting $\langle \hat{\beta}_R(\lambda), \hat{\beta}_R(\lambda) \rangle$ would be a better choice in this regard but it seems not to be done.) In addition, some authors have argued that the “exact” value of λ is not that important. But simulation studies have shown that even small changes in λ can produce rather large changes in MSE, so it is not clear that one can be so cavalier.

Example 9.4 We consider the Hald data to illustrate the ridge trace procedure for different biasing constants λ . From the original model ($m = 4, n = 13$), let us consider the regression in correlation form. That is, the new centered and scaled regressors z_j are

$$z_j = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}, \quad j = 1, 2, \dots, m, \quad i = 1, 2, \dots, n \quad (9.70)$$

where $S_{jj} = \sum_{i=1}^{13} (x_{ij} - \bar{x}_j)^2$, and take \mathbf{y} to be only centered. Then, the $m \times m$ matrix $\mathbf{Z}^T \mathbf{Z}$ is the correlation matrix of the \mathbf{x} 's, namely

$$\mathbf{Z}^T \mathbf{Z} = \begin{bmatrix} 1.00000 & 0.22858 & -0.82414 & -0.24545 \\ 0.22858 & 1.00000 & -0.13924 & -0.97296 \\ -0.82414 & -0.13924 & 1.00000 & 0.02954 \\ -0.24545 & -0.97296 & 0.02954 & 1.00000 \end{bmatrix},$$

the mean and the standard deviation for each predictor variable is

$$\begin{aligned} \bar{x}_1 &= 7.4615, & \bar{x}_2 &= 48.154, & \bar{x}_3 &= 11.769, & \bar{x}_4 &= 30.000, \\ s_1 &= 5.8824, & s_2 &= 15.561, & s_3 &= 6.4051, & s_4 &= 16.738. \end{aligned}$$

Also, the mean of \mathbf{y} , \bar{y} equals 95.423.

To obtain the ridge solution for the Hald data, we need to solve the equation

$$(\mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{I}) \hat{\boldsymbol{\beta}}_R = \mathbf{Z}^T \mathbf{y} \tag{9.71}$$

for various values of λ , $0 \leq \lambda \leq 1$. The ridge standardized coefficients for selected values of λ are presented in Table 9.1.

Table 9.1 Standardized Ridge Coefficients for λ				
λ	$\hat{\beta}_1^R(\lambda)$	$\hat{\beta}_2^R(\lambda)$	$\hat{\beta}_3^R(\lambda)$	$\hat{\beta}_4^R(\lambda)$
.000	31.6064	27.4984	2.2602	-8.3552
.002	28.7892	20.3721	-0.8305	-15.8571
.004	27.8876	18.3384	-1.7883	-17.9885
.006	27.3958	17.3886	-2.2900	-18.9766
.010	26.8001	16.5006	-2.8629	-19.8849
.020	25.9251	15.8039	-3.6179	-20.5432
.050	24.2839	15.5514	-4.8362	-20.5395
.100	22.4072	15.6238	-6.0297	-19.9286
.500	16.0618	14.6060	-8.0746	-16.2726
1.000	12.8080	12.6889	-7.5705	-13.5439

In (9.71) when $\lambda = 0$, we obtain the standardized least squares estimates of $\boldsymbol{\beta}$. In order to convert them to the original estimates $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m)$, we use the following conversion formula

$$\hat{\beta}_j = \hat{\beta}_j^R / \sqrt{S_{jj}}, \quad j = 1, 2, ..., m \tag{9.72}$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^m \hat{\beta}_j \bar{x}_j. \tag{9.73}$$

Using (9.72) we obtained the estimated ridge coefficients which are shown in Table 9.2. As we noted, the ridge coefficients for $\lambda = 0$ coincide with the regression coefficients in the standard fit. In order to examine the behavior of the ridge coefficients, a graph of the ridge trace is shown in Figure 9.2.

Table 9.2 Estimated Ridge Coefficients for λ				
λ	$\hat{\beta}_1^R(\lambda)$	$\hat{\beta}_2^R(\lambda)$	$\hat{\beta}_3^R(\lambda)$	$\hat{\beta}_4^R(\lambda)$
.000	1.55106	0.51013	0.10187	-0.14410
.002	1.41281	0.37793	-0.03743	-0.27348
.004	1.36857	0.34020	-0.08060	-0.31024
.006	1.34443	0.32258	-0.10320	-0.32728
.010	1.31520	0.30611	-0.12903	-0.34295
.020	1.27226	0.29318	-0.16306	-0.35430
.050	1.19172	0.28850	-0.21797	-0.35424
.100	1.09962	0.28984	-0.27176	-0.34370
.500	0.78822	0.27096	-0.36392	-0.28065
1.000	0.62854	0.23539	-0.34120	-0.23359

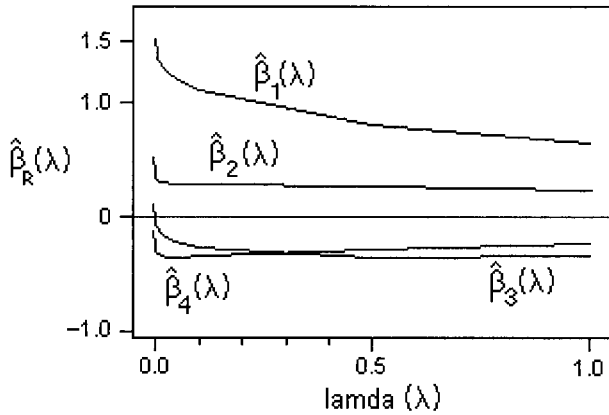


Figure 9.2: Ridge trace for the Hald data using four regressors

As we see from the graph and Table 9.1, the desirable value of λ can be chosen between 0.01 and 0.02.

The ridge trace is a very sensible and pragmatic way of choosing the *shrinkage parameter* λ . Since as λ gets bigger, the variance reduces, and the coefficients become more stable. A value for λ is chosen at the point for which the coefficients are no longer changing rapidly. However, it should be noted that stability does not imply that the regression coefficients have converged.

In addition to this, we can think of a couple of other plausible criteria to look at. One is the values of the VIF - values near 1 are desirable, and another is the coefficient of multiple determination R^2 for each value of λ .

(2) The Harmonic Mean Estimator A simple, objective estimator for λ was given by Hoerl, Kennard and Baldwin in [62]. They proposed using

$$\hat{\lambda} = ms^2 / \langle \hat{\beta}, \hat{\beta} \rangle \quad (9.74)$$

where $\hat{\beta}$ is the least squares estimator of β and s^2 is the usual estimate of variance in the model. This choice of λ can be motivated by considering the harmonic mean of the optimal ridge parameters in *generalized ridge regression*. (This derivation will be discussed in Section 9.5) and as measured by reduction in MSE was one of three best in the simulation study of ten different estimators by Galarneau-Gibbons [39], although a previous study by Gunst and Mason [46] had a large standard deviation and could be quite variable in shape. On the other hand, it is easy to calculate, has appealing statistical interpretations and requires no subjective judgements on the part of the analyst.

A possible improvement of this estimator was suggested in Hoerl and Kennard [61]. There λ was defined as the assumed limit of the sequence

$$\hat{\lambda}_{j+1} = ms^2 / \langle \hat{\beta}_R(\lambda_j), \hat{\beta}_R(\lambda_j) \rangle, \quad j \geq 0, \quad (9.75)$$

where λ_0 is given by (9.74). A complicated rule for terminating the iteration is given in [61] while Galarneau-Gibbons terminated the iterations when $|\lambda_{j+1} - \lambda_j| < 10^{-4}$ and

defaults to the least squares estimator if convergence is not obtained in 30 iterations. This rule also performed well in her study, but not always better than the noniterative version.

Example 9.5 From the previous Example 9.4, since the number of regressors in the model $m = 4$, the standard error from the standard fit $s = 2.446$ and the standardized least squares estimates of β from Table 9.1 are

$$\hat{\beta}_R^T(0) = (\beta_1, \beta_2, \beta_3, \beta_4) = (31.6064, 27.4984, 2.2602, -8.3552).$$

Using (9.74) a reasonable choice of λ is

$$\hat{\lambda} = ms^2 / (\hat{\beta}^T \hat{\beta}) = 4(2.446)^2 / 1830.044 \simeq 0.0130771.$$

Thus, when we choose $\hat{\lambda} = 0.0131$,

$$\hat{\beta}_R^T = (26.4778, 16.1652, -3.1528, -20.2158),$$

and using (9.72)-(9.73), we have the resulting fitted ridge regression model

$$\hat{y}_R = 83.414 + 1.30x_1 + 0.30x_2 - 0.1421x_3 - 0.3487x_4.$$

We note that in this model the sign of $\hat{\beta}_3$ is now negative, $\hat{\beta}_0$ is bigger, and both $\hat{\beta}_1$ and $\hat{\beta}_2$ are smaller than the ones in the original model.

(3) SRIDG This estimator of λ is based on the observation that the value of λ which minimizes $MSE(\hat{\beta}_R)$ is given as the solution to the equation (obtained by differentiating the MSE in (9.63) with respect to λ)

$$\sum_{i=1}^m \frac{\lambda \gamma_i^2 - \sigma^2}{(\lambda + \lambda_i)^3} = 0. \quad (9.76)$$

Estimating σ^2 in (9.68) by s^2 from a preliminary least squares fit and γ_i by

$$\hat{\gamma}_i(\lambda) = [\mathbf{Q}^T \hat{\beta}_R(\lambda)]_i, \quad 1 \leq i \leq m \quad (9.77)$$

an approximation $\hat{\lambda}$ to the value of λ satisfying (9.76) is obtained by evaluating

$$|s(\lambda)| = \left| \sum_{i=1}^m \frac{\lambda \hat{\gamma}_i(\lambda) - s^2}{(\lambda + \lambda_i)^3} \right| \quad (9.78)$$

for a range of values of λ and choosing $\hat{\lambda}$ as the value which minimizes $|s(\lambda)|$. This estimator was also one of the three best in [39].

(4) Bayes Estimators A number of stochastic estimators have been based on a Bayesian interpretation of the ridge estimator. In this context, we assume that each regression coefficient has a prior distribution which is $N(0, \sigma_\beta^2)$ and then the Bayes estimator $\hat{\beta}_B$ of β is

$$\hat{\beta}_B = (\mathbf{X}^T \mathbf{X} + \sigma^2 / \sigma_\beta^2)^{-1} \mathbf{X}^T \mathbf{y} \quad (9.79)$$

which may be interpreted as a ridge estimator with ridge parameter $\lambda = \sigma^2 / \sigma_\beta^2$. Replacing σ^2 and σ_β^2 by various estimators gives a whole series of methods for estimating λ .

For instance, if $\sigma_\beta^2 = E(\langle \beta, \beta \rangle)$, (remember now that β is a random vector) σ_β^2 is estimated by

$$\hat{\sigma}_\beta^2 = \frac{1}{m} \sum_{i=1}^m \hat{\beta}_i^2 = \frac{1}{m} \langle \hat{\beta}, \hat{\beta} \rangle \quad (9.80)$$

where $\hat{\beta}_i$'s are the OLS estimators of β_i , and s^2 is the usual estimate σ^2 then λ may be estimated by

$$\hat{\lambda} = \frac{s^2}{\hat{\sigma}_\beta^2} = \frac{ms^2}{\langle \hat{\beta}, \hat{\beta} \rangle} \quad (9.81)$$

which is just the harmonic mean estimator. A similar estimator, $\hat{\lambda} = (m+2)s^2 / \langle \hat{\beta}, \hat{\beta} \rangle$ and iterative version were proposed by Lindley and Smith in [75]. Further examples of estimators derived from a Bayesian point of view may be found in [84, 105].

(5) Constrained Estimators As we have noted previously, the ridge estimator may be viewed as a constrained least squares estimator with the goal of ridge estimation to produce "shorter" estimates of β than $\hat{\beta}$. Using this as motivation McDonald and Galarneau [79] developed a method for estimating λ based on (9.24) which shows that

$$E(\langle \hat{\beta}, \hat{\beta} \rangle) = \langle \beta, \beta \rangle + \sigma^2 \sum_{i=1}^m \frac{1}{\lambda_i}. \quad (9.82)$$

Estimating $E(\langle \hat{\beta}, \hat{\beta} \rangle)$ by $\langle \hat{\beta}_R, \hat{\beta}_R \rangle$ and β by $\hat{\beta}$, they suggest choosing λ so that

$$\langle \hat{\beta}_R, \hat{\beta}_R \rangle = \langle \hat{\beta}, \hat{\beta} \rangle - s^2 \sum_{i=1}^m \frac{1}{\lambda_i} \quad (9.83)$$

which will constrain $\langle \hat{\beta}_R, \hat{\beta}_R \rangle$ to be smaller than $\langle \hat{\beta}, \hat{\beta} \rangle$, provided that (9.82) has a solution. If (9.82) is not solvable, i.e., the right hand side is negative then they considered using $\lambda = 0$ (i.e., the least squares estimator or $\lambda = \infty$ and $\hat{\beta}_R = 0$). In their simulations [79] neither method proved better (in terms of MSE) than least squares in all cases.

(6) PRESS(λ) If improvement in prediction rather than fit is the primary goal of one's analysis, then methods which estimate λ based on prediction performance statistics can be used. A straightforward approach along these lines is to use a generalization of the PRESS statistic discussed in Chapter 8.

Letting $\hat{\beta}_{R(-i)}(\lambda)$ denote the ridge estimate of β with the i -th observation deleted and $\hat{y}_{(-i)}(\lambda) = \langle \hat{\beta}_{R(-i)}(\lambda), \mathbf{x}_i \rangle$ the corresponding predicted value at \mathbf{x}_i , then we define $\text{PRESS}(\lambda)$ by

$$\text{PRESS}(\lambda) = \sum_{i=1}^m [y_i - \hat{y}_{(-i)}(\lambda)]^2 \quad (9.84)$$

and note that $\text{PRESS}(0) = \text{PRESS}$. Choosing λ to minimize $\text{PRESS}(\lambda)$ then yields a ridge estimator, which hopefully has “better” prediction properties than $\hat{\beta}$. (However, keep in mind our comments concerning prediction in Section 9.5.)

Unfortunately, if $\lambda > 0$, no simple expression for $\text{PRESS}(\lambda)$ exists analogous to that for $\text{PRESS}(0)$. Thus a large amount of computation is necessary to evaluate this statistic and to search for the minimizing value of λ . An approximation to $\text{PRESS}(\lambda)$ which is cheaper to compute may be found in [90].

(7) C_λ Statistic Using a similar prediction optimization philosophy Mallows in [82] introduced a generalization of the C_p statistic as a tool for estimating λ . He proposed plotting

$$C_\lambda = \frac{\text{SSE}(\lambda)}{s^2} - n + 2 + 2[\text{tr}(\mathbf{X}\mathbf{L})] \quad (9.85)$$

where $\mathbf{L} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_m)^{-1}\mathbf{X}^T$, against

$$V_\lambda = 4 [\text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{L}\mathbf{L}^T)] \quad (9.86)$$

and choosing λ to minimize C_λ .

(8) Generalized Cross Validation Estimator Perhaps the most popular prediction based statistic for choosing λ is Wahba, Golub and Heath's *generalized cross validation* (GCV) statistic [116] which they motivate as a rotation invariant version of $\text{PRESS}(\lambda)$. It is defined by

$$\text{GCV}(\lambda) = \frac{\|(\mathbf{I}_m - \mathbf{H}(\lambda))\mathbf{y}\|^2}{[\text{tr}(\mathbf{I}_m - \mathbf{H}(\lambda))]^2} \quad (9.87)$$

where $\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_m)^{-1}\mathbf{X}^T$ is the ridge generalization of the hat matrix since $\hat{\mathbf{Y}}(\lambda) = \mathbf{H}(\lambda)\mathbf{y}$ gives the vector of predicted ridge estimates of \mathbf{Y} at the observation points \mathbf{x}_i , $1 \leq i \leq n$.

If $\lambda = 0$, then,

$$\text{GCV}(0) = \frac{\|(\mathbf{I}_m - \mathbf{H})\mathbf{y}\|^2}{[\text{tr}(\mathbf{I}_m - \mathbf{H})]^2} = \frac{\text{SSE}}{(n - \sum_{i=1}^m \tilde{h}_{ii})^2} = \frac{\text{SSE}}{(n - m)^2} = s^2 \quad (9.88)$$

where \tilde{h}_{ii} is the i -th diagonal element of $\mathbf{X}_{sc}(\mathbf{X}_{sc}^T\mathbf{X}_{sc})^{-1}\mathbf{X}_{sc}^T$. In this case we notice that $\text{GCV}(0) \neq \text{PRESS}(0)$ and for $\lambda = 0$, GCV is actually a measure of fit rather than of prediction.

As for $\text{PRESS}(\lambda)$ λ is chosen to minimize $\text{GCV}(\lambda)$. This statistic has fared well in simulations [39] and has been widely used in nonstatistical applications of ridge type procedures.

Nonstochastic Estimators

Nonstochastic estimators of λ appear to be less used in practice. As with the ridge trace, they are somewhat subjective. We mention only a few.

(1) Reduces VIFs If one examines the formula for $\Sigma(\hat{\beta}_R)$ it is easy to show that the variance inflation factors of the ridge estimators $\hat{\beta}_R(\lambda)$ are decreasing functions of λ . If improved estimation is one's goal, then a simple procedure is to choose λ so that all the VIFs are between one and ten. The smallest value of λ which does this is preferable, for then the bias is minimized as well (see [8]).

(2) $\text{tr}[\mathbf{H}(\lambda)]$ In [90] Myers proposes examining $\text{tr}[\mathbf{H}(\lambda)]$ which may be regarded as the effective number of regression degrees of freedom in the problem. From the expression $h_{ii}(\lambda_i)/(\lambda_i + \lambda)$ which are the diagonal elements of $\mathbf{H}(\lambda)$

$$DF(\lambda) = \text{tr}[\mathbf{H}(\lambda)] = \sum_{i=1}^m \frac{\lambda_i}{\lambda_i + \lambda}, \quad (9.89)$$

one sees that $DF(\lambda)$ is a decreasing function of λ with $DF(0) = m$. As with the ridge trace, one can generally expect a rapid fall off of $DF(\lambda)$ in a plot of $DF(\lambda)$ against λ and then stabilization.

The ridge parameter λ is chosen as the point where this stabilization occurs. The value of $DF(\lambda)$ where this occurs is referred to as the *effective rank* of \mathbf{X}_{sc} [112].

9.5.4 Generalized Ridge Regression

In ridge regression one is attempting to control the stability of the coefficient estimates through the use of a single parameter. Since there is no reason to believe that a single value of λ can stabilize all the estimates simultaneously, it is perhaps more reasonable to try to use m parameters to control each coefficient separately. This idea leads to the notion of *generalized ridge regression* which we discuss next.

We begin by examining the effect of ridge regression on the canonical form of the model and this will lead to a straightforward generalization. From (9.41)

$$\hat{\beta}_R = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{X} \hat{\beta} \quad (9.90)$$

and using the spectral decomposition of $\mathbf{X}^T \mathbf{X}$ this gives (using $\mathbf{Q}^T = \mathbf{Q}^{-1}$)

$$\hat{\beta}_R = (\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T + \lambda \mathbf{I}_m)^{-1} \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \hat{\beta} = \mathbf{Q} (\mathbf{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{\Lambda} \mathbf{Q}^T \hat{\beta} \quad (9.91)$$

so that

$$\mathbf{Q}^T \hat{\beta}_R = (\mathbf{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{\Lambda} \mathbf{Q}^T \hat{\beta}. \quad (9.92)$$

If we let $\hat{\gamma}_R = \mathbf{Q}^T \hat{\beta}_R$ and noting that $\mathbf{Q}^T \hat{\beta} = \hat{\gamma}$, the least squares estimator $\hat{\gamma}$ in the canonical form of the linear model (recall Example 5.11), then (9.92) becomes

$$\hat{\gamma}_R = (\mathbf{\Lambda} + \lambda \mathbf{I}_m)^{-1} \mathbf{\Lambda} \hat{\gamma} \quad (9.93)$$

so that

$$\hat{\gamma}_{R,i} = \frac{\lambda_i \hat{\gamma}_i}{\lambda + \lambda_i}. \quad (9.94)$$

Since $\text{Var}(\hat{\gamma}_i) = \sigma^2/\lambda_i$ and

$$\text{Var}(\hat{\gamma}_{R,i}) = \left(\frac{\lambda_i}{\lambda + \lambda_i} \right)^2 \frac{\sigma^2}{\lambda_i} = \frac{\lambda_i \sigma^2}{(\lambda + \lambda_i)^2}, \quad (9.95)$$

the effect of ridge regression is to reduce the variance of $\hat{\gamma}_i$ by a factor of $\lambda_i^2/(\lambda + \lambda_i)^2$ which decreases to zero as $\lambda \rightarrow \infty$. Thus, ridge regression counteracts the effects of small eigenvalues on the canonical estimates by multiplying the least squares estimators by a *filter* of the form $\lambda_j/(\lambda + \lambda_j)$. Because there is generally no a priori reason to use a common filter for each coefficient, we can consider replacing λ by a different value μ_i for each component and this leads one to define *generalized ridge estimators* for γ by $\hat{\gamma}_{GR,i}$ where

$$\gamma_{GR,i} = \frac{\lambda_i \hat{\gamma}_i}{\lambda_i + \mu_i}, \quad 1 \leq i \leq m. \quad (9.96)$$

The generalized ridge estimator for $\hat{\beta}$ is then given by

$$\hat{\beta}_{GR} = \mathbf{Q} \hat{\gamma}_{GR}. \quad (9.97)$$

As for ordinary ridge regression, we are now faced with the task of choosing μ_i , $1 \leq i \leq m$. As before, a reasonable approach is to determine these to minimize the mean square error of $\hat{\gamma}_{GR,i}$, $1 \leq i \leq m$. Now,

$$MSE(\hat{\gamma}_{GR,i}) = [E(\hat{\gamma}_{GR,i}) - \gamma_i]^2 + \text{Var}(\hat{\gamma}_{GR,i}) \quad (9.98)$$

and from (9.98) we find that $E(\hat{\gamma}_{GR,i}) = \lambda_i \gamma_i / (\lambda_i + \mu_i)$ (since $\hat{\gamma}_i$ is unbiased) and $\text{Var}(\hat{\gamma}_{GR,i}) = \lambda_i \sigma^2 / (\lambda_i + \mu_i)^2$ so that

$$MSE(\hat{\gamma}_{GR,i}) = \frac{(\mu_i \gamma_i)^2}{(\lambda_i + \mu_i)^2} + \frac{\sigma^2 \lambda_i}{(\lambda_i + \mu_i)^2} = \frac{(\mu_i \gamma_i)^2 + \sigma^2 \lambda_i}{(\lambda_i + \mu_i)^2}. \quad (9.99)$$

Using standard calculus arguments it can be shown that

$$\mu_i = \sigma^2 / \gamma_i^2, \quad (9.100)$$

minimizes (9.99). Since γ_i and σ^2 are unknown, these are typically replaced by their OLS estimates $\hat{\gamma}_i$ and s^2 leading to stochastic estimators of μ_i

$$\hat{\mu}_i = s^2 / \hat{\gamma}_i^2, \quad 1 \leq i \leq m. \quad (9.101)$$

These may then be used in place of μ_i in (9.100) and then β is estimated by (9.97).

An alternative approach to estimating μ_i is to attempt to minimize $MSE(\hat{\beta}_{GR})$. However, using calculations similar to those in Theorem 9.1.

$$MSE(\hat{\beta}_{GR}) = \sum_{i=1}^m MSE(\hat{\gamma}_{GR,i}) \quad (9.102)$$

and minimizing (9.102) leads to the same values of μ_i given in (9.100). Thus both approaches give the same result.

As for ordinary ridge regression iterative versions of the estimators $\hat{\mu}_i, 1 \leq i \leq n$, have been considered [58]. These are defined by the sequence

$$\begin{cases} \hat{\mu}_{i,j} = s^2 / \hat{\gamma}_{i,j}^2 \\ \hat{\gamma}_{i,j+1} = \lambda_i \hat{\gamma}_{i,j} / (\lambda_i + \hat{\mu}_{i,j}), \quad j = 0, 1, 2, \dots \end{cases} \quad (9.103)$$

where $\hat{\gamma}_{i,0}, 1 \leq i \leq m$, is the least squares estimator of γ_i . β_i is then estimated by the sequence

$$\hat{\beta}_{i,j} = (\mathbf{Q}\hat{\gamma}_j)_i, \quad i = 1, 2, \dots, m, \quad j = 0, 1, 2, \dots \quad (9.104)$$

where $\hat{\gamma}_i = (\hat{\gamma}_{1,j}, \hat{\gamma}_{2,j}, \dots, \hat{\gamma}_{m,j})^T$.

The iteration is terminated when the lengths of successive iterates $\hat{\gamma}_j, \hat{\gamma}_{j+1}$ are approximately the same. In this regard, it is interesting to note that Hemmerle in [54] gave a closed form solution for $\lim_{j \rightarrow \infty} \hat{\beta}_{i,j}$ so that carrying out the iterations in (9.103) is actually unnecessary.

In particular, let \mathbf{T} be the $m \times m$ diagonal matrix $\text{diag}(\tau_1, \tau_2, \dots, \tau_m)$ where

$$\tau_i = \begin{cases} 0, & \text{if } \hat{\gamma}_i^2 \lambda_i / s^2 < 4 \\ 1/2 + [1/4 - s^2 / \lambda_i \hat{\gamma}_i^2]^{1/2}, & \text{otherwise} \end{cases} \quad (9.105)$$

then

$$\lim_{j \rightarrow \infty} \hat{\gamma}_{i,j} = (\mathbf{T}\hat{\gamma})_i \quad (9.106)$$

and so

$$\lim_{j \rightarrow \infty} \hat{\beta}_{i,j} = (\mathbf{Q}\mathbf{T}\hat{\gamma})_i. \quad (9.107)$$

If we observe that $(\hat{\gamma}_i^2 \lambda_i / s^2)^{1/2}$ is the t -statistic associated with the least squares estimator $\hat{\gamma}_i$, then the fully iterated generalized ridge estimator (9.103) says to estimate γ_i by zero if the observed t statistic is not significant at about the 5% level, otherwise γ_i is estimated by the function of the least squares estimator given in (9.106). Thus, “nonsignificant” least squares estimates are shrunk to zero, while the remaining ones are shrunk less drastically.

It was observed by Hemmerle [54] that (9.103) often induced too much shrinkage (bias) in the estimates of γ_i and he and others have proposed various modifications of this procedure. (Further details may be found in [55].) We also note that these estimators were not evaluated in the simulation study of Gibbons [39].

As a further observation on generalized ridge regression we note, as stated in Section 9.5 that the harmonic mean estimator of λ in ordinary ridge regression may be viewed as the harmonic mean of the generalized ridge parameters $\hat{\mu}_i = s^2 / \hat{\gamma}_i^2$. For this, we observe that by definition, the harmonic mean of $\hat{\mu}_i, 1 \leq i \leq m$, is given by

$$\begin{aligned} \frac{m}{\sum_{i=1}^m (1/\hat{\mu}_i)} &= \frac{m}{\sum_{i=1}^m (\hat{\gamma}_i^2 / s^2)} = \frac{ms^2}{\sum_{i=1}^m \hat{\gamma}_i^2} \\ &= \frac{ms^2}{\langle \hat{\gamma}, \hat{\gamma} \rangle} = \frac{ms^2}{\langle \mathbf{Q}^T \hat{\beta}, \mathbf{Q}^T \hat{\beta} \rangle} = \frac{ms^2}{\langle \hat{\beta}, \hat{\beta} \rangle}. \end{aligned} \quad (9.108)$$

As a last comment on generalized ridge regression we observe that another choice of parameters is often made. That is, if $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$ (this is the reverse order from what we have been using), then one chooses $\mu_i = \infty$, $1 \leq i \leq r$, and $\mu_i = 0$, $r+1 \leq i \leq m$, where $\lambda_1, \lambda_2, \dots, \lambda_r$ are the r smallest eigenvalues of $\mathbf{X}^T \mathbf{X}$. Often, as in the study of Gunst and Mason [46], $r = 1$.

This gives rise to a class of generalized ridge estimators called *principal components estimators*. Denoting these estimators with the sub/superscript “pc” we have

$$\hat{\gamma}_i^{pc} = \begin{cases} 0, & i = 1, 2, \dots, r, \\ \hat{\gamma}_i, & i = r+1, r+2, \dots, m. \end{cases} \quad (9.109)$$

The corresponding estimator of β is given by

$$\hat{\beta}_{pc} = \mathbf{Q} \hat{\gamma}_{pc} \quad (9.110)$$

where $\hat{\gamma}^{pc} = (\hat{\gamma}_1^{pc}, \hat{\gamma}_2^{pc}, \dots, \hat{\gamma}_m^{pc})^T$.

Further discussion of these estimators will be given in the next section.

9.6 Other Alternatives to OLS

Ridge estimation is by no means the only alternative to OLS that has been proposed to improve estimation and prediction when multicollinearity is present. Here we discuss two further approaches, *mixed estimation* and *principal components regression*. A description of other methods, such as *latent root regression*, may be found in [51].

9.6.1 Mixed Estimation

This is a quasi-Bayesian approach for solving ill-conditioned problems, which uses an a priori constraint on β as additional phoney data to improve the quality of estimation.

Suppose then that we have a standard linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ and assume that we have a set of $p < m$ prior constraints on β which can be written in the form

$$\mathbf{z} = \mathbf{D}\beta + \delta \quad (9.111)$$

where for simplicity we have $E(\delta) = 0$ and $\Sigma(\delta) = \sigma_1^2 \mathbf{I}_p$, and, in general, $\sigma_1^2 \neq \sigma^2$. Then augmenting the basic model with this information we get

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{D} \end{pmatrix} \beta + \begin{pmatrix} \varepsilon \\ \delta \end{pmatrix}. \quad (9.112)$$

If ε and δ are assumed independent, then the mixed estimator $\hat{\beta}_M$ of β is the generalized least squares estimator obtained from (9.111) and is given by

$$\hat{\beta}_M = \left[\frac{\mathbf{X}^T \mathbf{X}}{\sigma^2} + \frac{\mathbf{D}^T \mathbf{D}}{\sigma_1^2} \right]^{-1} \left[\frac{\mathbf{X}^T \mathbf{y}}{\sigma^2} + \frac{\mathbf{D}^T \mathbf{z}}{\sigma_1^2} \right], \quad (9.113)$$

and observe that if $\mathbf{D}^T \mathbf{D} = \mathbf{I}_m$ and $\mathbf{z} = 0$ then $\hat{\beta}_M$ becomes the ridge estimator

$$\hat{\beta}_R = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_m)^{-1} \mathbf{X}^T \mathbf{y} \quad (9.114)$$

where $\lambda = \sigma^2/\sigma_1^2$ and is identical to the normal theory Bayes estimator. Again this result emphasizes that ridge estimation is OLS with constraints imposed on β via prior information.

A numerical example, taken from economic theory may be found in BKW [8].

9.6.2 Principal Components Estimation

As we pointed out in our discussion of generalized ridge regression, the ridge parameters can be chosen to make the coefficients $\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_r$ ($\hat{\gamma}_i$ is the least squares estimate of γ_i in the canonical model) corresponding to the r smallest eigenvalues of $\mathbf{X}^T \mathbf{X}$ equal to zero. Doing this leads to the principal components estimators.

$$\hat{\beta}_{pc}(r) = \mathbf{Q} \hat{\gamma}_{pc}(r), \quad r = 1, 2, \dots, m \quad (9.115)$$

where

$$\hat{\gamma}_{pc}(r) = (0, 0, \dots, \hat{\gamma}_{r+1}, \hat{\gamma}_{r+2}, \dots, \hat{\gamma}_m)^T \quad (9.116)$$

where the eigenvalues of $\mathbf{X}^T \mathbf{X}$ have been ordered as $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$ with the columns of \mathbf{Q} being ordered in a corresponding fashion.

For $1 \leq r \leq m$, $\hat{\beta}_{pc}$ is generally biased with

$$E(\hat{\beta}_{pc}) = \mathbf{Q}_2 \gamma_2 \quad (9.117)$$

where $\hat{\gamma}$ is partitioned so that $\gamma = [\gamma_1 | \gamma_2]^T$, where $\gamma_1 = (\gamma_1, \dots, \gamma_r)^T$ and $\gamma_2 = (\gamma_{r+1}, \dots, \gamma_m)^T$ and \mathbf{Q} is partitioned as $[\mathbf{Q}_1 | \mathbf{Q}_2]$ to conform with this so that the columns of \mathbf{Q}_2 are the eigenvectors corresponding to $\lambda_{r+1}, \dots, \lambda_m$.

Thus, the bias is given by

$$\text{Bias} = \beta - \mathbf{Q}_2 \gamma_2 = \mathbf{Q}_1 \gamma_1 \quad (9.118)$$

and $\Sigma[\hat{\beta}_{pc}(r)]$ is given by

$$\Sigma[\hat{\beta}_{pc}(r)] = \mathbf{Q}_2 \Sigma(\hat{\gamma}_2) \mathbf{Q}_2^T = \sigma^2 \mathbf{Q}_2 \Lambda_2^{-1} \mathbf{Q}_2^T \quad (9.119)$$

where $\Lambda_2 = \text{diag}(\lambda_{r+1}, \dots, \lambda_m)$.

From (9.117) and (9.118) or (9.119)

$$MSE[\hat{\beta}_{pc}(r)] = \sum_{j=1}^r \gamma_j^2 + \sum_{j=r+1}^m \frac{\sigma^2}{\lambda_j}. \quad (9.120)$$

In order to make PC estimation operational, a choice of r must be made which is analogous to selecting the ridge parameter(s) in ridge regression to minimize MSE. That this is theoretically possible comes from examining (9.120) where it is seen that the bias component increases as r increases, whereas the variance component decreases. However, since $\gamma_i, 1 \leq i \leq m$, are unknown in practice, as for ridge regression, this choice of r cannot be made a priori. Because of this, a number of different selection philosophies are currently used.

One approach is to use variable selection techniques, such as adding components starting with $r = m - 1$ until some measure of fit such as R^2 stabilizes. Another approach is to use *backward elimination* starting with the full model and eliminating γ_i 's on the basis of the standard t -tests. This, latter approach was criticized by Gunst and Mason [46] as a consequence of the poor showing of PC estimators this procedure gave in the simulation study of Dempster, Schatzoff and Wermuth in [26]. They use (but not clearly recommend) the PC estimator with $r = 1$. That is, they eliminate the component corresponding to the smallest eigenvalue λ_1 . This seems sensible since it can be shown that the principal components estimator $\hat{\beta}_{pc}(r)$ is the least squares estimator of β constrained by setting the dependencies determined by the r -th smallest eigenvalues to zero. So, $\hat{\beta}_{pc}(1)$ is determined by minimizing $\langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle$ subject to

$$\sum_{i=1}^m q_i \beta_i = 0 \quad (9.121)$$

where $(q_1, q_2, \dots, q_n)^T$ is the eigenvector for $\lambda_{\min}(\mathbf{X}^T \mathbf{X})$.

In their simulation studies they found, as measured by MSE, that the PC(1) estimator generally outperformed ordinary least squares and ridge regression with the harmonic mean estimator (9.75) for the ridge parameter, for ill-conditioned problems, but was worse than least squares for nearly orthogonal data.

Further details may be found in [26]. However, because their simulations were all performed with $\langle \beta, \beta \rangle = 1$ the same criticism as given by Draper and Van Nostrand [28] for ridge regression simulations holds.

9.7 Exercises

9.1 Is ridge regression a least squares procedure?

9.2 Consider the hosing price data in Example 5.12

- (a) Find the sample correlations between the regressor variables.
- (b) What are the variance inflation factors?
- (c) Find the condition number of $\mathbf{X}^T \mathbf{X}$. Give a comment.

9.3 Using the birth weight data in Example 5.15, answer the followings.

- (a) Find the sample correlations matrix among regressor variables.
- (b) Calculate the variance inflation factors.
- (c) Find the eigenvalues of $\mathbf{X}^T \mathbf{X}$.
- (d) Find the condition number of $\mathbf{X}^T \mathbf{X}$. Give a comment.

9.4 Consider the Longley data in Example 5.17.

- (a) Find the sample correlation matrix among the regressor variables.
- (b) Using the numbers in (a), give a brief comment about the indication of multicollinearity.
- (c) What are the variance inflation factors?

- (d) Find the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$.
- (e) Find the condition number of $\mathbf{X}^T\mathbf{X}$. Give a comment for the evidence of multicollinearity.
- (f) Find the ridge regression solution for the data.

9.5 A researcher obtained the following data from an experiment.

Obs. No.	Y	x_1	x_2
1	39	0	0
2	44	0	0
3	45	0	0
4	42	1	-1
5	38	-1	1
6	35	-1	-1
7	37	1	1

- (a) Fit the data to the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$.
- (b) Can you suggest a different model(s) to fit the data?

9.6 Consider the data below for the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Obs. No	Y	x_1	x_2
1	10	-2	-4
2	6	-1	-1
3	6	0	0
4	12	1	1
5	14	2	4
Sum	48	0	0
Sum of Squares	512	10	34

- (a) Center and scale the x_1 , x_2 and Y columns.
- (b) Write out the normal equations for (a) and solve them.
- (c) Find the determinant of the correlation matrix.

9.7 Prove Theorem 9.1 (vi).

9.8 Prove that the ridge estimator is the solution to the problem

$$\min_{\beta} (\beta - \hat{\beta})^T \mathbf{X}^T \mathbf{X} (\beta - \hat{\beta}) \text{ subject to } \beta^T \beta \leq d^2.$$

9.9 Stein [109] proposed the pure shrinkage estimator defined as $\hat{\beta}_s = c\hat{\beta}$, where $0 \leq c \leq 1$, and c is a constant chosen by the experimenter. Show that the pure shrinkage estimator is the solution to the problem

$$\min_{\beta} (\beta - \hat{\beta})^T (\beta - \hat{\beta}) \text{ subject to } \beta^T \beta \leq d^2.$$

9.10 Consider the following data set given in below.

y	x_1	x_2	x_3	x_4
11.7	10.0	12125	132.2	404.6
17.9	11.5	36717	501.5	1180.6
21.1	11.6	43319	904.0	1807.5
14.7	11.2	10530	227.6	470.0
7.7	10.7	3931	66.6	151.4
8.4	10.0	1536	43.4	93.8
32.8	6.8	61400	1253.0	3293.4
17.6	8.8	2589	83.1	158.2
10.9	8.5	1186	24.2	96.2
9.2	7.7	291	4.5	31.8
16.2	4.9	1276	9.1	95.0
10.1	9.6	6633	158.2	407.2

- (a) Find the eigenvalues and eigenvectors of scaled \mathbf{X} matrix (not centered).
- (b) What is the condition number of $\mathbf{X}^T \mathbf{X}$? Give a comment for the evidence of multicollinearity.
- (c) Find the variance inflation factors for the regression coefficients.
- (d) Draw the ridge trace for the data and find the ridge regression solution.
- (e) Do you have any suggestions to alleviate the multicollinearity?

Appendix

Statistical Tables

Table A.1	Standard Normal Distribution
Table A.2	Percentiles of the Student's t -Distribution
Table A.3	Percentiles of the Chi-Square Distribution
Table A.4a	F -Distribution - Critical values of upper 10% points
Table A.4b	F -Distribution - Critical values of upper 5% points
Table A.4c	F -Distribution - Critical values of upper 1% points
Table A.5	Critical Values of the Durbin-Watson Statistic

Table A.2 Percentiles of the Student's t -Distribution

$$P(T \leq t) = \int_{-\infty}^t \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r} \Gamma(r/2) (1+x^2/r)^{(r+1)/2}} dx, \quad P(T \leq -t) = 1 - P(T \leq t)$$

Note: Entry is the value of t . For lower percentiles, use the relation $t_{\alpha} = -t_{1-\alpha}$. In particular, $t_{.50} = -t_{.50} = 0$. For example, for degrees of freedom $r = 8$, $t_{.95} = -t_{.05} = 1.860$.

d.f. r	$P(T \leq t_p) = p$								
	0.75	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	127.322	318.309	636.619
2	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.684	1.315	1.706	2.056	2.479	2.799	3.067	3.435	3.707
27	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	.680	1.299	1.676	2.009	2.403	2.678	2.938	3.261	3.496
60	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Table A.3 Percentiles of the Chi-Square Distribution

$$P(X \leq x) = \int_{-\infty}^x \frac{1}{\Gamma(r/2)2^{r/2}} u^{(r/2)-1} e^{-u/2} du, \quad P(X \leq \chi_p^2) = p$$

Note: Entry is the value of x ($= \chi_p^2$). Notation: $.0^4 = .0000$

d.f. r	$P(X \leq \chi_p^2) = p$							
	0.005	0.025	0.05	0.90	0.95	0.975	0.99	0.995
1	.0 ⁴ 393	.0 ³ 982	.0 ² 393	2.706	3.841	5.024	6.635	7.879
2	.0100	.0506	.103	4.605	5.991	7.378	9.210	10.597
3	.0717	.216	.352	6.251	7.815	9.348	11.345	12.838
4	.207	.484	.711	7.779	9.488	11.143	13.277	14.860
5	.412	.831	1.145	9.236	11.070	12.832	15.086	16.750
6	.676	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	.989	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	3.940	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	4.575	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	5.226	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	5.892	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	6.571	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	7.261	22.307	24.996	27.488	30.578	32.801
16	5.142	6.908	7.962	23.542	26.296	28.845	32.000	34.267
17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401
22	8.643	10.982	12.338	30.813	33.924	36.781	40.289	42.796
23	9.260	11.689	13.091	32.007	35.172	38.076	41.638	44.181
24	9.886	12.401	13.848	33.196	36.415	39.364	42.980	45.558
25	10.520	13.120	14.611	34.382	37.652	40.646	44.314	46.928
26	11.160	13.844	15.379	35.563	38.885	41.923	45.642	48.290
27	11.808	14.573	16.151	36.741	40.113	43.194	46.963	49.645
28	12.461	15.308	16.928	37.916	41.337	44.461	48.278	50.993
29	13.121	16.047	17.708	39.087	42.557	45.722	49.588	52.336
30	13.787	16.791	18.493	40.256	43.773	46.979	50.892	53.672
35	17.192	20.569	22.465	46.059	49.802	53.203	57.342	60.275
40	20.707	24.433	26.509	51.805	55.758	59.342	63.691	66.766
50	27.991	32.357	34.764	63.167	67.505	71.420	76.154	79.490
60	35.535	40.482	43.188	74.397	79.082	83.298	88.379	91.952
70	43.275	48.758	51.739	85.527	90.531	95.023	100.425	104.215
80	51.172	57.153	60.391	96.578	101.879	106.629	112.329	116.321
90	59.196	65.647	69.126	107.565	113.145	118.136	124.116	128.299
100	67.328	74.222	77.929	118.498	124.342	129.561	135.807	140.169

Source: [73] Abridged from Table 9 of Kokoska, S. and Nevison, C. (1989), *Statistical Tables and Formulae*, Springer-Verlag, New York.

Table A.4a F-Distribution - Critical values of upper 10% points

$$P(X \leq f) = \int_0^f \frac{\Gamma[(v_1+v_2)/2] (v_1/v_2)^{v_1/2} u^{v_1/2-1}}{\Gamma(v_1/2)\Gamma(v_2/2)(1+v_1u/v_2)^{(v_1+v_2)/2}} du, \quad P(F > f_{v_1, v_2; 1-\alpha}) = \alpha = .10$$

Note: $F = s_1^2/s_2^2 = [S_1/v_1]/[S_2/v_2]$ where s_1^2 and s_2^2 are independent mean squares estimating common variance σ^2 and based on v_1 and v_2 degrees of freedom, respectively.

d.f. v_2	v_1 = Degrees of Freedom for Numerator													
	1	2	3	4	5	6	7	8	9	10	12	15	20	30
1	39.86	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.3
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.82
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.16
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.08
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.96
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.91
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.84
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.81
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.78
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.76
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.72
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.70
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.69
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.67
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.66
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.65
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.64
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.63
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.62
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.54
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.48
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.41
∞	2.17	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.34

Source: [94] Abridged from Pearson, E. and Hartley, H. (1966), *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., Cambridge University Press, Cambridge.

Table A.4b F-Distribution - Critical values of upper 5% points

$$P(X \leq f) = \int_0^f \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2} u^{v_1/2-1}}{\Gamma(v_1/2)\Gamma(v_2/2)(1+v_1u/v_2)^{(v_1+v_2)/2}} du,$$

$$P(F > f_{v_1, v_2; 1-\alpha}) = \alpha = .05$$

Note: $F = s_1^2/s_2^2 = [S_1/v_1]/[S_2/v_2]$ where s_1^2 and s_2^2 are independent mean squares estimating common variance σ^2 and based on v_1 and v_2 degrees of freedom, respectively.

d.f. v_2	v_1 = Degrees of Freedom for Numerator													
	1	2	3	4	5	6	7	8	9	10	12	15	20	30
1	161.4	200	216	225	230	234	237	239	241	242	244	246	248	250
2	18.51	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.41	19.4	19.4	19.5
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.01
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	1.96
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.88
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.85
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46

Source: [94] Abridged from Pearson, E. and Hartley, H. (1966), *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., Cambridge University Press, Cambridge.

Table A.4c F-Distribution - Critical values of upper 1% points

$$P(X \leq f) = \int_0^f \frac{\Gamma[(v_1+v_2)/2] (v_1/v_2)^{v_1/2} u^{v_1/2-1}}{\Gamma(v_1/2)\Gamma(v_2/2)(1+v_1u/v_2)^{(v_1+v_2)/2}} du, \quad P(F > f_{v_1, v_2; 1-\alpha}) = \alpha = .01$$

Note: $F = s_1^2/s_2^2 = [S_1/v_1]/[S_2/v_2]$ where s_1^2 and s_2^2 are independent mean squares estimating common variance σ^2 and based on v_1 and v_2 degrees of freedom, respectively.

d.f. v_2	v_1 = Degrees of Freedom for Numerator													
	1	2	3	4	5	6	7	8	9	10	12	15	20	30
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6261
2	98.50	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.42	99.4	99.5	99.5
3	34.12	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.4	27.2	27.05	26.9	26.7	26.5
4	21.20	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.6	14.37	14.2	14.0	13.8
5	16.26	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.38
6	13.75	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.72
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.62
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.47
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.41
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70

Source: [94] Abridged from Pearson, E. and Hartley, H. (1966), *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., Cambridge University Press, Cambridge.

Table A.5 Critical Values of the Durbin-Watson Statistic

This table provides the two limiting values of critical d (d_L and d_U), corresponding to the two most extreme configurations of the regressors for testing autocorrelation. Note that the critical values are one-sided. (Significance level α = probability in lower tail.)

For example, suppose there are $n = 20$ observations and $p = 3$ regressors, and we wished to test $H_0: \rho = 0$ versus $H_1: \rho > 0$ at $\alpha = .05$. Then if D fell below $d_L = 1.00$, we would reject H_0 . If D were above $d_U = 1.68$, we could not reject H_0 . If D were between d_L and d_U , our decision is indecisive.

Sample Size n	α	p = Number of Independent Variables (Excluding the Constant)									
		1		2		3		4		5	
		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	.01	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
	.025	.95	1.23	.83	1.40	.71	1.61	.59	1.84	.48	2.09
	.05	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
16	.01	.84	1.09	.74	1.25	.63	1.44	.53	1.66	.44	1.90
	.025	.98	1.24	.86	1.40	.75	1.59	.64	1.80	.53	2.03
	.05	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15
18	.01	.90	1.12	.80	1.26	.71	1.42	.61	1.60	.52	1.80
	.025	1.03	1.26	.93	1.40	.82	1.56	.72	1.74	.62	1.93
	.05	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06
20	.01	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74
	.025	1.08	1.28	.99	1.41	.89	1.55	.79	1.70	.70	1.87
	.05	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
25	.01	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
	.025	1.18	1.34	1.10	1.43	1.02	1.54	.94	1.65	.86	1.77
	.05	1.29	1.45	1.21	1.55	1.12	1.66	1.01	1.77	.95	1.89
30	.01	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
	.025	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	.98	1.73
	.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	.01	1.25	1.34	1.20	1.40	1.25	1.46	1.10	1.52	1.05	1.58
	.025	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
	.05	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
	.025	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
	.05	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
	.025	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
	.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
80	.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
	.025	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
	.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
100	.01	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65
	.025	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72
	.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Source: [30] Abridged from Tables I, II and III of Durbin, J. and Watson, G. (1951), *Biometrika*, Vol. 38, pp. 159-177.

Bibliography

- [1] Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets. *Technometrics*, **16** 221-227.
- [2] Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13** 469-475.
- [3] Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, **16** 523-531.
- [4] Anscombe, F. J. and Tucky, J. W. (1963). The examination and analysis of residuals. *Technometrics*, **5** 141-160.
- [5] Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables. *Journal of the Royal Statistical Society, Series B*, **44** 1-36.
- [6] Atkinson, A. C. (1983). Diagnostic regression for shifted power transformations. *Technometrics*, **25** 23-33.
- [7] Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20** 1-6.
- [8] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, Inc., New York.
- [9] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26** 211-243.
- [10] Box, G. E., Hunter, W. G. and Hunter, J. S. (1978). Statistics for Experimenters. John Wiley & Sons, Inc., New York.
- [11] Box, G. E., Jenkins, G. and Reinsel, G. (1994). Time Series Analysis. 3rd ed., Prentice Hall, New Jersey.
- [12] Box, G. E. and Tidwell, P. W. (1962). Transformation of the independent variables, *Technometrics*, **4** 531-550.
- [13] Casella, G. and Berger, R. (2002). Statistical Inference. 2nd ed., Duxbury, Pacific Grove, California.

- [14] Charnes, A., Frome, E. and Yu, P. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, **71** 169-171.
- [15] Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*, John Wiley & Sons, Inc., New York.
- [16] Chatfield, C. (1970). Discrete distributions in market research. *Random Counts in Physical Science, Geosciences, and Business* (G. P. Patil, ed.), Pennsylvania University Press, University Park, Pennsylvania.
- [17] Cochran, D. and Orcutt, G. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, **44** 32-61.
- [18] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19** 15-18.
- [19] Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, **74** 169-174.
- [20] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.
- [21] Copas, J. B. (1983). Regression prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society, Series B*, **45** 311-354.
- [22] Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*. 2nd ed., John Wiley & Sons, Inc., New York.
- [23] David, M. (1977). *Geostatistical Ore Reserve Estimation*. Elsevier Scientific Publishing Co., Amsterdam.
- [24] DeLury, D. B. (1960). *Values and Integrals of the Orthogonal Polynomials up to $N = 26$* . University of Toronto, Toronto.
- [25] Dobson, A. J. (1990). *Introduction to Generalized Linear Models*. Chapman and Hall, London.
- [26] Dempster, A., Schatzoff, M. and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, **72** 77-90.
- [27] Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. 3rd ed., John Wiley & Sons, Inc., New York.
- [28] Draper, N. R. and Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimation: review and comments. *Technometrics*, **21** 451-466.
- [29] Durbin, J. and Watson, G. (1950). Testing for serial correlation in least squares regression I, *Biometrika*, **37** 409-438.
- [30] Durbin, J. and Watson, G. (1951). Testing for serial correlation in least squares regression II, *Biometrika*, **38** 159-177.

- [31] Durbin, J. and Watson, G. (1971). Testing for serial correlation in least squares regression III, *Biometrika*, **58** 1-19.
- [32] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - An empirical Bayes approach. *Journal of the American Statistical Association*, **68** 117-130.
- [33] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70** 311-319.
- [34] Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, **17** 111-117.
- [35] Forsythe, A., Engelman, L., Jennrich, R., and May, P. (1973). A stopping rule for variable selection in multiple regression. *Journal of the American Statistical Association*, **68** 75-77.
- [36] Franke, R. (1982). Scattered data interpolation: Tests of some methods. *Mathematics of Computation*, **38** 181-200.
- [37] Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, **37** 152-155.
- [38] Gibbons, D. G. (1979). A Simulation Study of Some Ridge Estimators. General Motors Research Laboratories, Mathematics Department, GMR-2659 (rev. ed.), Warren, Michigan.
- [39] Gibbons, D. G. (1981). A Simulation Study of Some Ridge Estimators. *Journal of the American Statistical Association*, **76** 131-139.
- [40] Golberg, M. A. (1984). An Introduction to Probability Theory with Statistical Applications. Plenum Press, New York.
- [41] Golberg, M. A. and Chen, C. S. (1997). Discrete Projection Methods for Integral Equations, Computational Mechanics Publications, Southampton.
- [42] Golberg, M. A. and Chen, C. S. (1994). The theory of radial basis functions applied to the BEM for inhomogeneous partial differential equations. *Boundary Elements Communications*, **5** 57-61.
- [43] Golberg, M. A. and Chen, C. S. (1998). The method of fundamental solutions for potential, Helmholtz and diffusion problems. Boundary Integral Methods: Numerical and mathematical aspects, 103-176. Editor: Golberg, M., WIT Press, Southampton.
- [44] Gorman, J. and Toman, R. (1966). Selection of variables for fitting equations to data. *Technometrics*, **8** 27-51.
- [45] Graybill F. A. (1976). Theory and Application of the Linear Model. Wadsworth & Brooks/Cole, Pacific Grove, California.
- [46] Gunst, R. F. and Mason, R. L. (1977). Biased estimation in regression: An evaluation using mean squared error, *Journal of the American Statistical Association*, **72** 616-628.

- [47] Hahn, G. (1972). Simultaneous prediction intervals for a regression model. *Technometrics*, **14** 203-214.
- [48] Hahn, G. (1973). The coefficient of determination exposed! *Chemical Technology*, **3** 609-614.
- [49] Hald, A. (1952). Statistical Theory with Engineering Applications. John Wiley & Sons, Inc., New York.
- [50] Hardy, R. L. (1990). Theory and applications of the multiquadric biharmonic method. *Computational Mathematics and Applications*, **19** 163-208.
- [51] Hawkins, D. (1973). On the investigation of alternative regressions by principal components analysis. *Applied Statistics*, **22** 275-286.
- [52] Healy, M. J. (1990). GLIM: An introduction. Oxford University Press. Oxford.
- [53] Helms, R. W. (1974). The average estimated variance criterion for the selection-of-variables problem in general linear models. *Technometrics*, **16** 261-273.
- [54] Hemmerle, W. (1975). An explicit solution for generalized ridge regression. *Technometrics*, **17** 309-314.
- [55] Hemmerle, W. and Brantle, T. (1978). Explicit and constrained generalized ridge regression. *Technometrics*, **20** 109-120.
- [56] Hocking, R. R. (1974). Misspecification in regression. *American Statistician*, **28** 39-40.
- [57] Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32** 1-49.
- [58] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12** 55-67.
- [59] Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems, *Technometrics*, **12** 69-82.
- [60] Hoerl, A. E. and Kennard, R. W. (1975). A note on a power generalization of ridge regression. *Technometrics*, **17** 269.
- [61] Hoerl, A. E. and Kennard, R. W. (1976). Ridge regression: Iterative estimation of the biasing parameter, *Communications in Statistics*, **A5** 77-88.
- [62] Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: Some simulations, *Communications in Statistics*, **4** 105-123.
- [63] Hogg, R. and Craig, A. (1995). Introduction to Mathematical Statistics. 5th ed., Prentice Hall, New Jersey.
- [64] Hosmer, D. W. and Lemeshow, S. (2000). Applied logistic regression. 2nd ed., John Wiley & Sons, Inc., New York.
- [65] Huber, P. (1981). Robust Statistics. John Wiley & Sons, Inc., New York.

- [66] Huber, P. (1996). Robust Statistical Procedures. 2nd ed., SIAM, Philadelphia.
- [67] James, W. and Stein, C. (1961). Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California, Berkeley.
- [68] Kempthorne, O. (1957) An Introduction to Genetic Statistics. John Wiley & Sons, Inc., New York.
- [69] Kendall, M. and Stuart, A. (1979). The Advanced Theory of Statistics, Volume II: Inference and Relationships, 4th ed., Macmillan, New York.
- [70] Kendall, M. and Yule, G. (1950). An Introduction to the Theory of Statistics, Charles Griffin, London.
- [71] Kennard, R. W. (1971). A note on the C_p statistic. *Technometrics*, **13** 899-900.
- [72] Kmenta, J. (1986). Elements of Econometrics. 2nd., Macmillan, New York.
- [73] Kokoska, S. and Nevison, C. (1989). Statistical Tables and Formulae, Springer-Verlag, New York.
- [74] Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation. 2nd ed., Springer-Verlag, New York.
- [75] Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **34** 1-18.
- [76] Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point view of the user. *Journal of the American Statistical Association*, **62** 819-841.
- [77] Looney, S. and Gullledge, T. Jr. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician*, **39** 75-79.
- [78] Madych, W. (1992). Miscellaneous error bounds for multiquadric and realed interpolants. *Computational Mathematics and Applications* **24** 121-138.
- [79] McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, **70** 407-416.
- [80] McNamara, A. and Browne, J. (1980). An interesting correlation - United States oil and gold prices. *The New York Statistician*, **32** 2.
- [81] Mason, R. L., Gunst, R. F. and Webster, J. T. (1975). Regression analysis and problems of multicollinearity. *Communications in Statistics*, **4**(3) 277-292.
- [82] Mallows. C. L. (1973). Some comments on C_p . *Technometrics*, **15** 661-675.
- [83] Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, **12** 621-625.
- [84] Marquardt, D. and Snee, R. (1975). Ridge regression in practice. *The American Statistician*, **12** 3-19.

- [85] McCullagh, P. and Nelder, J. (1983). Generalized linear models. Chapman and Hall, London.
- [86] Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician*, **38** 124-126.
- [87] Montgomery, D., Peck, E. and Vining, G. (2001). Introduction to Linear Regression Analysis. 3rd ed., John Wiley & Sons, Inc., New York.
- [88] Montgomery, D. (2001). Design and Analysis of Experiments. 5th ed., John Wiley & Sons, Inc., New York.
- [89] Mood, A., Graybill, F. and Boes, D. (1963). Introduction to the Theory of Statistics. McGraw-Hill, Inc. New York.
- [90] Myers, R. H. (1990). Classical and modern regression with applications. 2nd ed., Brooks/Cole, Pacific Grove, California.
- [91] Narula, S. C. (1974). Predictive mean square error and stochastic regressor variables. *Applied Statistics*, **23** 11-16.
- [92] Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135** 370-384.
- [93] Neter, J., Wasserman, W. and Kutner, M. (1985). Applied Linear Regression Models, Irwin Inc., Chicago.
- [94] Pearson, E. and Hartley, H. (1966). Biometrika Tables for Statisticians, Vol. 1, 3rd ed., Cambridge University Press, Cambridge.
- [95] Poggio, T. and Girosi, F. (1990). Networks for approximation and learning, *Proceedings of the IEEE*, **78** 1481-1497.
- [96] Powell, M. J. D. (1992). The theory of radial basis function approximation in 1990. *Advances in Numerical Analysis*, Vol. II, ed. W. Light, Oxford Sciences Publications, Oxford.
- [97] Rao, C. R. (1973). Linear Statistical Inference and Its Applications. 2nd ed., John Wiley & Sons, Inc., New York.
- [98] Rencher, A. and Pun, F. (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22** 49-53.
- [99] Ryan, D., Joiner, B. and Ryan Jr., T. (1985). MINITAB Handbook. 2nd ed., Duxbury, Boston.
- [100] Saha, A. and Wu, C. L. (1993). Approximation, dimension reduction and non convex optimization using linear superpositions of Gaussians, *IEEE Trans. On Computers*, **42** 1222-1233.
- [101] Scheffé, H. (1959). The Analysis of Variance. John Wiley & Sons, Inc., New York.
- [102] Schmidt, P. (1973). Calculating the power of the minimum standard error choice criterion. *International Economic Review*, **14**.

- [103] Searle, S. R. (1982). *Matrix Algebra useful for Statistics*. John Wiley & Sons, Inc., New York.
- [104] Seber, G. (1977). *Linear Regression Analysis*. John Wiley & Sons, Inc., New York.
- [105] Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods (with discussion). *Journal of the American Statistical Association*, **75** 74-103.
- [106] Smith, P. (1979). Splines as a useful and convenient statistical tool. *American Statistician*, **33**(2) 59-61.
- [107] Spjøtvoll, E. (1972a). On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics*, **43** 398-411.
- [108] Spjøtvoll, E. (1972b). Multiple comparison of regression functions. *The Annals of Mathematical Statistics*, **43** 1076-1088.
- [109] Stein, C. (1960). *Multiple regression. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, Stanford, California.
- [110] Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge, Massachusetts.
- [111] Stone, M. (1974). Cross-validating choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, **36** 117-147.
- [112] Strang, G. (1988). *Linear Algebra and its Applications*. 3rd ed., Brooks and Cole. Pacific Grove, California.
- [113] Thompson, M. L. (1978a). Selection of variables in multiple regression: Part I. A review and evaluation. *Inst. Stat. Rev.*, **46** 103-106.
- [114] Thompson, M. L. (1978b). Selection of variables in multiple regression: Part II. A review and evaluation. *Inst. Stat. Rev.*, **46** 129-146.
- [115] Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.
- [116] Whaba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
- [117] Whaba, G., Golub, G. and Heath, C. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21** 206-211.
- [118] Weisberg, S. (1985). *Applied Linear Regression*. 3rd ed., John Wiley & Sons, Inc., New York.
- [119] Wendland, H. (1994). *Ein Beitrag zur interpolation mit radialen basis funktionen*, Diplomarbeit, Universität Göttingen, Göttingen, Germany.
- [120] Wetherill, G. B. (1986). *Regression Analysis with Applications*. Chapman and Hall, London.

- [121] Wilm, H. G. (1950). Statistical control in hydrologic forecasting. *Res. Note* **61**, Pacific Northwest Forest Range Experiment Station, Oregon.
- [122] Wold, S. (1974). Spline functions in data analysis. *Technometrics*, **16** 1-11.
- [123] Wonnacott, R. and Wonnacott, T. (1981). Regression: A second course in Statistics. John Wiley & Sons, Inc., New York.
- [124] Wu, Z. (1994). Multivariate compactly supported positive definite radial functions, *Technical Report*, Univerisitat Göttingen, Göttingen, Germany.

Index

- Adjusted R-square, 224
- Alias, 361
- Analysis of covariance model, 182
- Analysis of variance, 82
 - general linear model, 224
 - lack of fit analysis, 106
 - m-way, 332
 - table, 86
 - two-way, 332
- Analysis of variance model, 182
- Andrew's test, 282, 286
- Assessment function, 361
- Atkinson's modification, 278
- Augmented coefficient matrix, 164
- Autocorrelation
 - Cochrane-Orcutt method, 302
 - coefficient, 301
 - detecting, 301
 - Durbin-Watson statistic, 302
 - Hildreth-Lu procedure, 304
- Auxillary regressions, 390
- Average estimated variance, 367
- Backward elimination, 369, 371, 409
- Backward substitution, 162
- Basis
 - orthonormal, 148
- Bayes rule, 31
- Bayes' theorem, 6
- Best linear unbiased estimator, 55, 217
- Best test, 41
- Beta function, 24, 32
- Beta weights, 195
- Bias, 33, 48
- Bias matrix, 361
- Biased estimator, 389
- binary 0-1 predictors, 341
- binary 0-1 responses, 341
- Binomial coefficient, 9
- Bivariate normal distribution, 178
- BLUE estimator, 217
- Bonferroni inequality, 78
- Box-Cox method, 299
- Box-Cox transformation, 119
- Box-Tidwell method, 118, 277
- Canonical form, 196, 350
- Cauchy-Schwarz inequality, 143, 144, 302, 310
- Centered observations, 191
- Centering matrix, 173
- Central limit theorem, 22, 25, 41, 76
- Characteristic polynomial, 150
- Chi-square distribution, 22
- Cholesky factorization, 166, 188, 307
- Cochrane-Orcutt method, 302
- Coefficient matrix, 161
 - augmented, 164
- Coefficient of determination, 83
 - adjusted, 362
- Complex conjugate, 150
- Condition index, 388
- Condition number, 169, 382
 - ill-conditioned, 170
 - optimally conditioned, 170
 - well-conditioned, 170
- Conditional Poisson distribution, 47
- Confidence interval, 37
 - approximate, 39
 - exact, 37
 - lower limit, 45
 - one-sided interval, 45
- Consistency, 33
- Convergence
 - in probability, 33
- Cook's distance, 272, 285
- Correlation coefficient, 17
 - sample, 83

- Correlation form, 195, 214
- Covariance, 16
- Cramer's rule, 75, 137, 193, 212
- Criteria functions, 361
- Cross-validation, 326
- Cumulative distribution function, 8
 - joint, 12
 - properties, 10
- Data
 - asymmetric, 25
- Degrees of freedom, 22, 23
 - denominator, 24
 - numerator, 24
- Deletion statistics, 266
- Density, 11
 - conditional, 14
 - joint discrete, 13
 - posterior, 31
- Dependent variable, 51
- Design matrix, 167, 183, 352
 - centered, 191
 - centered and scaled, 194
 - linearly dependent, 335
- Design variable, 51
- Determinant, 137
- Deviance, 352
- DFBETAS, 272
- DFFITS, 273
- Diagonalization, 149
- Distribution
 - Bernoulli, 9
 - binomial, 9
 - chi-square, 22
 - double exponential, 55
 - F, 24
 - Laplace, 55
 - logistic, 11
 - lognormal, 25
 - normal, 21
 - of rare events, 9
 - Poisson, 9
 - conditional, 47
 - standard normal, 19
 - t, 23
 - uniform, 10
- Distribution function, 8
- Dot product, 143
- Double exponential distribution, 55
- Dummy variable, 181, 328
 - 0-1 coding, 335
 - analysis of variance model, 182
 - levels, 335
 - linearly dependent, 335
- Durbin-Watson statistic, 301
 - test procedure, 302
- Effective rank, 404
- Efficiency, 34
- Eigenvalue, 149
- Eigenvector, 149
- Error
 - heteroscedastic, 52
 - homoscedastic, 52
- Error vector, 183
- Errors in variable (EIV) model, 55
- Estimable model, 336
- Estimation
 - Bayesian, 31
 - bias, 48
 - least squares, 52
 - maximum likelihood, 25, 52
- Estimator
 - Bayes, 33
 - best, 36
 - best linear unbiased, 55
 - bias, 33
 - least squares, 31, 54, 74
 - linear, 73
 - maximum likelihood, 29
 - method of moments, 26
 - minimum variance unbiased, 34, 52
 - minimum variance unbiased linear, 73
 - properties
 - consistent, 33
 - efficient, 34
 - sufficient, 35
 - unbiased, 27, 33
- Euclidean norm, 323
- Events, 5
 - complement event, 6
 - independent events, 7
 - mutually exclusive events, 5
- Excluded category, 330
- Expectation, 15

- linearity properties, 16
- Exponential family, 12, 46, 350, 358
 - exponential density, 12
- Extra sum of squares principle, 238, 307
 - regression sum of squares, 238
- Extrapolation
 - hidden, 243
- F-distribution, 24
- F-test, 234
 - derivation of the F-test, 236
 - full model, 234
 - reduced model, 234
- Factoring, 165
- Factorization theorem, 35
- Filter, 405
- First order autoregressive, 301
- Forward selection, 369, 372
- Forward substitution, 162
- Full rank model, 188
- Fundamental theorem of algebra, 150
- Fundamental theorem of calculus, 11
- Gamma function, 22
- Gauss-Markov theorem, 55, 73, 217, 295
- Gauss-Newton method, 327
- Gaussian elimination, 137, 162
 - augmented coefficient matrix, 164
 - backward substitution, 162
 - factoring, 165
 - forward substitution, 162
 - partial pivoting, 163
 - pivot element, 163
 - round-off error, 163
- General linear model, 179
- Generalized cross validation, 403
- Generalized least squares
 - estimator, 306
 - weighted mean, 308
- Generalized likelihood ratio test, 43, 352
- Generalized linear model, 348
 - deviance, 352
 - generalized likelihood ratio test, 352
 - linear predictor, 349
 - link function, 349
 - weighted least squares, 352
- Generalized ridge regression, 404
 - estimators, 405
 - filter, 405
- Generation matrix, 173
- Geometric mean, 122
- Gram-Schmidt orthogonalization, 147
- Gram-Schmidt process, 168, 196
- Growth rate, 99
- Hat matrix, 207, 250
- Heteroscedasticity, 52
- Hildreth-Lu procedure, 304
- Homoscedasticity, 52
- Hypothesis
 - composite alternative, 40, 43
 - simple, 40
 - simple alternative, 40
- Hypothesis testing, 40
 - alternative hypothesis, 40
 - best test, 41
 - critical region, 40
 - Neyman-Pearson lemma, 41
 - null hypothesis, 40
 - significance level, 41
 - size of Type I error, 40
 - Type I and II errors, 40
- I Charts, 113
- i.i.d. random variable, 15
- Ill-conditioning, 167
- Independence, 14
- Independent variable, 51
- Independent variable hull (IVH), 243
- Inequality
 - Cauchy-Schwarz, 143
 - triangular, 144
- Influence diagnostics, 271
 - Cook's distance, 273
 - DFBETAS, 272
 - DFFITs, 273
- Inner product, 143
 - properties, 143
- Interaction term, 204
- Interactions, 337
 - two-way, 337
- Interpolation, 325
 - noisy data, 323
 - non-noisy data, 323
- Interval estimation, 37
- Iteration, 27, 352

- James-Stein estimator, 391
- Joint confidence region, 78, 80, 239
- Kriging, 301, 323, 325
- Lack of fit test, 104
 - bias, 104
 - lack of fit sum of squares, 105
 - pure error sum of squares, 105
- Lagrange multipliers, 74, 236, 247
- Laplace distribution, 55
- Laten root regression, 407
- Law of total probability, 6
- Least squares, 52
 - equations, 187
 - estimation, 31
 - estimator, 54
 - function, 244
 - ordinary, 188
 - orthogonal, 56
 - weighted, 288
- Least squares line, 54
 - intercept, 54
 - residual, 54
 - slope, 54
- Leverage, 251
 - high, 251
 - high leverage point, 251
- Likelihood function, 29, 120
 - log, 29
- Likelihood ratio, 41
- Likelihood ratio test
 - generalized, 43, 352
 - likelihood ratio, 41
- Linear estimator, 217
- Linear predictor, 349
- Linear transformations, 135
 - linear operator, 136
 - linearity, 136
 - mapping, 135
- Link function, 349
 - canonical, 350
 - identity, 349
 - log-log link, 349
 - power family, 350
 - probit, 349
- Logistic coefficients, 345
- Logistic distribution
 - cdf of, 342
- Logistic regression, 279
- Mallow's C_p , 363
 - properties, 365
- Marginal distribution, 13
- Matrix, 129
 - block form, 139
 - coefficient, 161
 - columns, 129
 - determinant, 137
 - diagonal, 130
 - diagonalization, 149
 - dimension, 130
 - elements, 130
 - diagonal, 130
 - off-diagonal, 130
 - full rank, 138
 - hat, 207
 - idempotent, 134
 - identity, 130
 - inverse, 136
 - non-symmetric, 130
 - orthogonal, 148
 - orthogonal projection, 135
 - partitioned form, 139
 - positive definite, 154
 - positive semidefinite, 154
 - power, 134
 - projection, 134
 - rank, 138
 - rectangular, 130
 - rows, 129
 - square, 130
 - submatrix, 139
 - sum, 131
 - symmetric, 130
 - trace, 135
 - transpose, 130
 - triangular
 - lower, 130
 - upper, 130
- Matrix addition, 131
 - associativity, 131
 - commutativity, 131
 - conformable, 131
- Matrix inversion, 136
- Matrix multiplication, 132

- associative property, 133
- coefficient matrix, 132
- commute, 133
- conformable, 133
- distributive, 133
- product, 132
- Maximum likelihood estimate, 184
- Maximum likelihood estimation, 25, 28, 52
- Mean square, 24
- Mean square error, 34
- Mean value, 16
- Median, 25
- Method of moments, 26
- Minimum absolute deviation line, 127
- Minimum variance unbiased estimator, 34, 52, 217
- Mixed estimation, 407
- Model misspecification, 359
- Model selection, 367
 - all possible regressions, 368
 - backward selection, 371
 - forward selection, 372
 - stepwise regression, 373
- Moment generating function, 18
- Moments
 - central, 16
 - n-th, 16
- Multicollinearity, 189, 227, 376
- Multiple linear model, 179
 - analysis of covariance, 182
 - analysis of variance, 182
 - design matrix, 183
 - dummy variable, 181
 - error random variable, 179
 - error vector, 183
 - independent variables, 179
 - least squares function, 244
 - linearity property, 181
 - normal equations, 211, 244
 - polynomial model of degree m , 181
 - quadratic model, 181
 - regression coefficients, 180
 - vector of observations, 183
 - vector of regression coefficients, 183
- Multivariate normal distribution
 - degenerate, 160
 - nondegenerate, 156
- Neural network modeling, 323
- Neyman-Pearson lemma, 41
- Nondecreasing function, 11
- Nonlinear estimator, 389
- Nonlinear regression, 327
- Normal equations, 187
- Normal plots, 252
- Normal probability plots, 111
- Numerical method
 - iteration, 27
 - Newton's, 351
 - total least squares, 56
- Odds ratio, 346
 - log odds, 346
 - odds, 346
- Ordinary least squares estimate, 188
- Orthogonal
 - basis, 146
 - hyperplane, 146
 - projection, 145
 - vector, 145
- Orthogonal design, 196
- Orthogonal least squares, 56
- Orthogonal matrix, 148
 - properties, 148
- Orthogonal polynomials, 316
- Orthonormal, 148
- Overfitting, 224
- Parallelogram law, 176
- Parameter space, 26
- Partial F-tests, 235
- Partial plots, 253
 - added variable plots, 253
 - regression plots, 253
 - residual plots, 253
- Partial regression coefficient, 180
- Penalized method, 365
- Piecewise polynomials, 277
- Pivot elements, 163
- Pivotal quantity, 37
- Pivoting, 163
- Polynomial models, 313, 323
 - multivariate, 322
 - orthogonal, 315
 - piecewise (or spline), 313
- Polynomial regression model, 181

- Positive definite, 154
- Positive semidefinite, 154
- Posterior distribution, 31
- Posterior mean, 32
- Posterior probability, 6
- Power family, 120, 350
- Power transformation, 299
- Powers of matrices, 134
- Prediction
 - confidence interval, 96
 - error sum of squares, 265
 - for a new observation, 95
 - for the mean response, 95
 - interval, 241
 - prediction interval, 96
 - standard error, 95
 - variance, 240
- PRESS residuals, 110
- Principal components
 - estimator, 407, 408
 - regression, 407
- Principal components form, 196
- Prior distribution, 31
- Prior probability, 6
- Probability, 5
 - conditional, 6
 - of an event, 5
 - posterior, 6
 - prior, 6
- Probability density function, 8
- Probability mass function, 8
- Probability space, 5
- Probit function, 342
- Probit link, 349
- Proper value, 174
- Pythagorean theorem, 144
- QR decomposition, 170, 188, 196, 379
- Quadratic form, 156
- Qualitative variable, 327
- Quantitative variable, 327
- R-square
 - adjusted, 224
- Radial basis functions, 323
 - Duchon splines, 324
 - Gaussian, 324
 - higher-order polyharmonic, 324
 - multiquadrics, 324
 - shape parameter, 324
 - thin-plate spline (TPS), 324
 - variance parameter, 324
- Random noise, 227
- Random sample, 15
- Random variable, 7
 - Bernoulli, 9
 - binomial, 9
 - canonical, 11
 - continuous random variable, 7
 - discrete, 7
 - independent, 14
 - logistic, 11
 - lognormal, 25
 - Poisson, 9
 - range of, 7
 - standard normal, 22
 - uniform, 10
- RBF approximations, 325
 - interpolation, 325
 - linear and nonlinear least squares, 325
 - smoothing interpolation, 325
- Regression coefficients, 180
 - partial, 180
- Regression function, 322
- Regression model
 - orthogonality
 - canonical form, 196
 - principal components form, 196
- Regression surface, 183
- Regression through the origin, 91
- Regularization, 394
- Relative risk, 346
- Residual, 54
 - PRESS, 265
 - studentized
 - externally, 251
 - internally, 251
- Residual plots, 297
 - added variable plots, 262
 - I Chart, 113, 256
 - normal plots, 252
 - partial plots, 253
 - partial regression plots, 262
 - variable plots, 252
- Residual sum of squares, 186

- standardized, 367
- Residuals
 - externally studentized, 110
 - outliers, 112
 - PRESS, 110
 - properties, 108
 - standardized, 110
 - studentized, 110, 112
- Residuals plots, 112
 - histogram, 111
 - normal probability plots, 111
- Response surface methodology, 323
 - response function, 322
 - response surface, 323
- Response variable, 51
- Ridge parameter, 326
- Ridge regression, 326, 376, 379, 380
 - Bayes estimators, 402
 - constrained estimators, 402
 - estimate, 393
 - estimated ridge coefficients, 399
 - estimators, 392
 - generalized, 400
 - harmonic mean estimator, 400
 - nonstochastic estimators, 404
 - parameter, 393
 - PRESS, 402
 - ridge trace, 398
 - shrinkage parameter, 400
 - SRIDG, 401
- Ridge trace, 398
- Robust regression analysis, 73
- Robustness, 38
- Round-off error, 163
- Sample correlations, 195
- Sample proportion, 39
- Sample space, 5
- Sample variance, 23, 27
- Scalar multiplication, 134
- Seasonality, 333
- Second derivative test, 30
- Selecting the best subset, 359
- Selection function, 361
- Serial correlation, 112
- Sherman-Morrison-Woodbury formula, 141
- Shrinkage parameter, 400
- Shunken estimator, 390
- shrinking factors, 391
- Significance level, 41
- Simple linear regression, 51
 - error model, 52
- Simultaneous confidence intervals, 78
- Singular matrix, 185
- Singular value decomposition, 167, 188, 380
 - singular vectors, 169
- Smoothing parameter, 326
- Spectral theorem, 149, 151, 210
- Spline functions, 317
- Spline models, 313, 323
- Splines, 277
- Standard deviation, 16
- Standard error, 65, 72
- Standard error of prediction, 95
- Standard normal distribution, 19
- Stepwise regression, 369, 373
- Stepwise regression variable selection, 229
- Studentized residual
 - externally, 251
 - internally, 251
- Sufficiency, 35
 - jointly sufficient, 36
- Sufficient statistic, 35
- Sum of squares
 - regression, 84
 - residual, 84
 - total, 84
- t-distribution, 23
- Test
 - generalized likelihood ratio, 43
 - likelihood ratio, 41
 - one-sided hypotheses, 45
 - two-tailed, 42
 - uniformly most powerful, 43
- Time series analysis, 301
- Time series data, 333
- Total least squares, 56
- Total variance, 214
- Trace, 135
- Transformations, 113, 276
 - Atkinson's modification, 278
 - Box-Cox method, 119, 280, 281
 - Box-Tidwell method, 118, 277, 281
 - in x , 276

- intrinsically linear, 118
 - linearizable, 280
 - logarithmic, 25
 - of y , 279
 - power family, 120
 - variance equalizing, 288
 - variance stabilizing, 299
- Tukey's rule, 367
- Two-tailed test, 44
- Unbiased estimator, 23
- Unbiasedness, 33
- Uncorrelatedness, 16
- Uniform measure, 10
- Uniformly most powerful test, 43
- Variable
- dependent, 51
 - design, 51
 - dummy, 328
 - independent, 51
 - qualitative, 327
 - quantitative, 327
 - response, 51
- Variable plots, 252
- Variable selection problem, 359
- Variance, 16, 17
- Variance decomposition proportion, 387
- Variance equalizing transformations
- weighted least squares, 288
- Variance inflation factor (VIF), 214
- Variance multiplication factor, 72, 77
- Variance stabilizing transformations, 299
- Variance-covariance matrix, 156
- Vector
- angle, 145
 - basis, 146
 - canonical base, 146
 - column space spanned, 146
 - dimension, 146
 - dot product, 143
 - inner product, 143
 - length, 144
 - linear combination, 138
 - linearly dependent, 138
 - linearly independent, 138
 - orthogonal, 145
 - orthogonal base, 146
 - orthonormal, 148
 - span, 146
 - subspace, 146
- Vector differentiation, 186, 243
- Vector of observations, 183
- centered, 191
- Vector of regression coefficients, 183
- Weighted least squares, 288
- equations, 352
 - estimator, 291
 - iteratively reweighted, 294
 - sum of squares, 290
 - weights, 289
- Weighted mean, 308
- Zero intercept model, 94